

Efficiency at Scale

CHROMA-1 for Morphological Profiling of Cell Painting Data

Benjamin Midtvedt¹, Jesús Pineda¹, Henrik Klein Moberg¹
Giovanni Volpe¹, Mattias Goksör¹

¹IFLAI AB

Abstract

The value of a Cell Painting screen depends on the biological information recoverable from its images and the replication needed to recover it. CHROMA-1 improves both. Without post hoc batch correction, it achieves 28% higher biological retrieval performance than the strongest tested alternative, even when each alternative receives its best-performing tested correction. This performance lead persists in substantially out-of-distribution settings, including changes to both staining protocols and cell lines. On RxRx3-core, CHROMA-1 matches or surpasses both Phenom-1 and Phenom-2 across five biological relationship endpoints despite training in a different cellular and imaging domain. These advantages lead to improved prediction of pharmacological activities. Across treatment-well budgets, CHROMA-1 matches CellProfiler’s macro predictive performance with approximately 50% fewer replicate wells for protein activity, mechanism of action and safety pharmacology, and achieves higher predictive performance when both use the same number of wells. CHROMA-1 also reveals substantial predictive signal for dose-qualified Ames mutagenicity in the same images, where CellProfiler remains close to random. Together, these results let image collections answer more biological questions while reducing the replication needed for useful pharmacological predictions.

Key results

- **28% higher cross-laboratory retrieval** without post hoc correction than any tested alternative with correction (Figure 1).
- **Highest scores on all five RxRx3-core endpoints** after zero-shot transfer; Phenom-1 and Phenom-2 trained on collections containing RxRx3 (Table 2).
- **Matches CellProfiler’s macro AP with about 50% fewer treatment wells** across budgets on assay activity prediction tasks (Figure 4).
- **One negative-control well per plate** preserves macro AP to within 0.1–0.4 percentage points on average across treatment budgets.
- **11.6 training GPU-hours; 0.87 inference GPU-hours** for 25,329 evaluation wells (Table 1; Figure 1B).

1 Introduction

Morphological profiling offers a way to ask many biological questions of the same experiment. By imaging multiple cellular compartments, Cell Painting captures a broad record of cellular

responses to chemical and genetic perturbations [1, 2]. CellProfiler has provided an established analytical foundation for this approach, converting images into quantitative measurements of shape, intensity and texture [3]. Beyond any single analysis, the opportunity is to make each screen a reusable source of biological knowledge whose value extends beyond the question that originally motivated it.

Learned representations have yet to consistently outperform CellProfiler across biological tasks and experimental settings. Recent benchmark preprints, including MorphoHELM [4] and JUMP-lite [5], show that CellProfiler remains a leading general-purpose baseline despite advances in microscopy-specific self-supervision and biologically guided vision models. Closing this gap requires more than higher benchmark scores: a reusable morphological profile must preserve biological relationships across changes in experimental conditions. Technical variation can obscure perturbation effects, and the removal of batch effects does not, by itself, establish that the relevant biology has been preserved [6]. The challenge is therefore to improve biological performance and retain that improvement beyond the training conditions.

In practice, the value of morphological profiling also depends on how much replication is needed to obtain an informative profile for each compound and condition. Every replicate well occupies capacity that could instead test another compound, concentration, or condition. Full-budget scores alone do not reveal how much replication is needed to reach a given level of predictive performance. Extracting more information from each treatment well could expand the scope of a screen within the same experimental budget. Computational demands also affect practicality. The proprietary Phenom models demonstrate strong biological retrieval [7], but their reported final training runs required 20,000 A100 GPU-hours for Phenom-1 and 48,000 H100 GPU-hours for Phenom-2 [8, 9]. Their model sizes of 307 million and 1.86 billion parameters, respectively, also impose substantial memory and computational requirements during inference [8, 9]. A useful representation should combine stronger biological performance with lower experimental and computational requirements.

Here, we introduce CHROMA-1, a state-of-the-art morphological profiling model, trained on a curated subset of CPG0016 COMPOUNDS and CRISPR plates [10, 11]. CHROMA-1 uses same-plate negative controls during inference to condition the representations on typical vehicle-control morphology. In the cross-batch retrieval benchmark proposed by Arevalo et al. [6], CHROMA-1 without post hoc batch correction achieves 28% higher non-replicate retrieval performance than the strongest tested alternative, single-cell MorphEM with Seurat CCA, even when that alternative uses its best-performing tested batch correction. Notably, this advantage persists across four out-of-domain evaluation settings spanning new acquisition sites, different staining protocols, and different cellular systems, including A549 cells and primary human hepatocytes. On RxRx3-core [7], CHROMA-1 leads on all five biological relationship endpoints, including against Phenom models trained on collections containing RxRx3, despite transferring from a different imaging domain and cell line.

Importantly, CHROMA-1’s advantages extend from biological retrieval to pharmacological prediction, where it matches CellProfiler’s macro predictive performance with approximately 50% fewer treatment wells across the evaluated budgets for protein activity, mechanism of action and safety pharmacology on BBBC047. When both methods use the full well budget, CHROMA-1 also achieves higher predictive performance. Beyond these established pharmacological readouts, CHROMA-1 reveals substantial predictive signal for dose-qualified Ames mutagenicity where the CellProfiler baseline remains close to random. These capabilities are obtained with a 50.4-million-parameter encoder and 11.6 GPU-hours of microscopy training (Table 1). Together, these results support extracting more biological information from existing image libraries and designing future screens to address more biological questions within the same well budget.

CHROMA-1 model overview

CHROMA-1 is a proprietary, 50.4-million-parameter convolutional encoder derived from ConvNeXt-Small [12]. It was trained on five-channel Cell Painting images from a curated subset of CPG0016 COMPOUNDS and CRISPR, using a cross-batch contrastive objective. The final run used 352×352 -pixel crops for 16,000 optimization steps on two GH200 accelerators, totaling 11.6 GPU-hours. At inference, the encoder processes complete fields rescaled to $0.65 \mu\text{m}$ per pixel, producing 768-dimensional site profiles that are aggregated into well profiles.

Same-plate vehicle controls provide the reference morphology during profile generation. The resulting profiles can be used directly, without fitting a separate batch-correction transform over the image collection.

Table 1: **CHROMA-1 training scale, model size, and inference throughput.** Training uses 352×352 -pixel crops; inference uses complete five-channel fields with 964–1,108 pixels per side on one GH200.

Encoder family	ConvNeXt-Small derived
Training crop size	352×352 pixels
Training steps	16,000
Profile dimensions	768
Inference field size	964–1,108 pixels per side
Training multichannel sites	2.8M
Encoder parameters	50.4M
Training GPU-hours	11.6
Training hardware	2× GH200
Inference sites/GPU-second	66.6

2 Raw CHROMA-1 beats batch-corrected alternatives on batch robustness

Reusing morphological profiles across screens requires comparing biological responses measured in different laboratories and on different microscopes. Technical differences between experiments can obscure these responses, making batch robustness essential for assembling useful image libraries. The aim is to preserve biological relationships across experiments, and removing technical variation is valuable only insofar as it supports that aim. We evaluate batch robustness and biological retrieval using the multi-laboratory, multi-microscope comparison described by Arevalo et al., following its scoring protocol. [6], which uses the TARGET2 plates of CPG0016. Biological retrieval is measured by nonreplicate mean average precision (mAP), which quantifies how well profiles retrieve wells treated with the same compound against wells treated with other compounds. We select mAP as our biological summary because the subsequent framework of Kalinin, Arevalo et al. establishes it as a single, sensitive measure of profile strength and similarity, with advantages over alternative metrics in high-dimensional profiling settings [13]. Batch mixing is summarized by the mean of four published metrics: batch silhouette [14, 15], graph connectivity [15], kBET [16] and batch LISI [17], following the convention established by Arevalo et al. [6]. Together, these assess separation by batch, connectivity among profiles of the same compound, and the composition of local neighborhoods. Higher scores indicate better performance on both axes.

We compare CHROMA-1 with CellProfiler and five learned representations. We include microscopy foundation models with public checkpoints that can profile multi-cell images: CellCLIP [18], OpenPhenom [19], and SubCell [20] (via MorphoHELM’s multi-cell adaptation [4]),

and MorphEM [21] (via JUMP-lite’s multi-cell adaptation). We also evaluated MorphEM as a single-cell model, due to its strong performance in JUMP-lite [5]. Phenom-1 and Phenom-2 have no public checkpoints, so we use their reported scores where available.

Without post hoc batch correction, CHROMA-1 achieves 28% higher nonreplicate mAP than the strongest tested alternative, single-cell MorphEM [21] with Seurat CCA [22] (Figure 1A). Single-cell MorphEM exceeds the strongest CellProfiler result, while the other learned models trail it. Applying Seurat CCA further improves CHROMA-1’s retrieval and batch mixing, increasing its retrieval advantage over the same comparator from 28% to approximately 34%. CHROMA-1’s separation represents an advance over both established morphological measurements and the evaluated learned representations.

CHROMA-1 also has the lowest measured profile-generation cost among the learned models (Figure 1B). It profiles the complete 25,329-well evaluation in 0.87 GH200 GPU-hours; the other learned models require 1.73–149.34 GPU-hours under their tested inference protocols. Notably, the strongest comparator, MorphEM single-cell, required 149.34 GPU-hours, approximately 171 times CHROMA-1’s recorded inference compute, excluding the cost of cell segmentation. The strongest cross-batch biological retrieval in this comparison therefore comes with the fastest profile generation and no separate post hoc correction step.

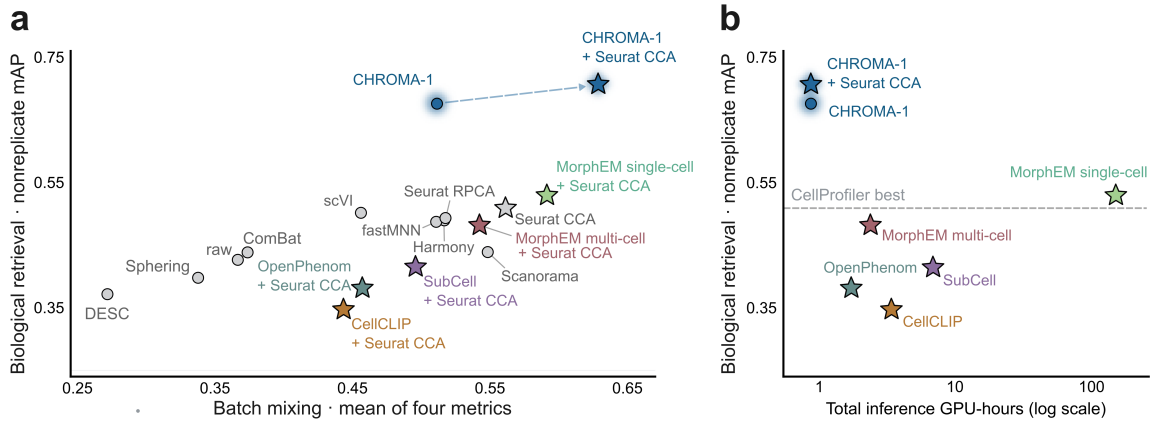


Figure 1: CHROMA-1 without post hoc correction outperforms the corrected alternatives in biological retrieval. (A) Biological retrieval versus batch mixing. CHROMA-1 is shown without post hoc correction and with Seurat CCA. Each other learned model is shown at its highest nonreplicate mAP across all tested correction methods; CellProfiler’s ten available conditions are shown in gray. Batch mixing is the mean of four benchmark metrics. (B) The same retrieval scores versus total recorded inference GPU-hours on GH200s, shown on a logarithmic axis. Recorded time includes image loading, preprocessing, and embedding inference, with post hoc correction timed separately. The dashed line marks the best CellProfiler retrieval score.

3 Leading retrieval performance extends across unseen experimental contexts

Strong retrieval across batches does not establish that a representation will remain useful as experiments depart further from its training conditions. To measure robustness in out-of-distribution scenarios, we evaluated CHROMA-1 along an escalating ladder of experimental differences. JUMP gene overexpression [23] introduces a new acquisition site and perturbation modality (Figure 2A). BBBC047 [24, 25] adds an independently acquired compound screen with

a different staining protocol (Figure 2B). LINCS [26] and OASIS [27] extend the evaluation to different cellular systems: A549 cells and primary human hepatocytes, respectively (Figure 2C). OASIS represents a particularly pronounced visual departure from the U2OS training images, as illustrated by the representative fields. Throughout, we follow Arevalo et al.’s nonreplicate mAP scoring procedure and CellProfiler feature preprocessing [6, 13]. Comparisons are made within each dataset because the retrieval candidate sets differ.

Without post hoc batch correction, CHROMA-1 leads every included alternative across all four evaluated settings, even when each comparator is represented by its best-performing tested processing condition. On JUMP gene overexpression, it achieves a nonreplicate mAP of 0.261, compared to 0.186 for CellProfiler, the strongest alternative (Figure 2A). On BBBC047, CHROMA-1 reaches 0.186 versus 0.103 for CellProfiler, the closest competitor that did not train on the dataset (Figure 2B). The advantage extends to A549 cells in LINCS, where CHROMA-1 reaches 0.667 versus 0.543, and to primary human hepatocytes in OASIS, where it reaches 0.123 versus 0.100 (Figure 2C). BBBC047 was included in CellCLIP’s training [18] and is therefore not an unseen dataset for that model. Single-cell MorphEM requires cell segmentations, which were unavailable for these collections, so it is not included here.

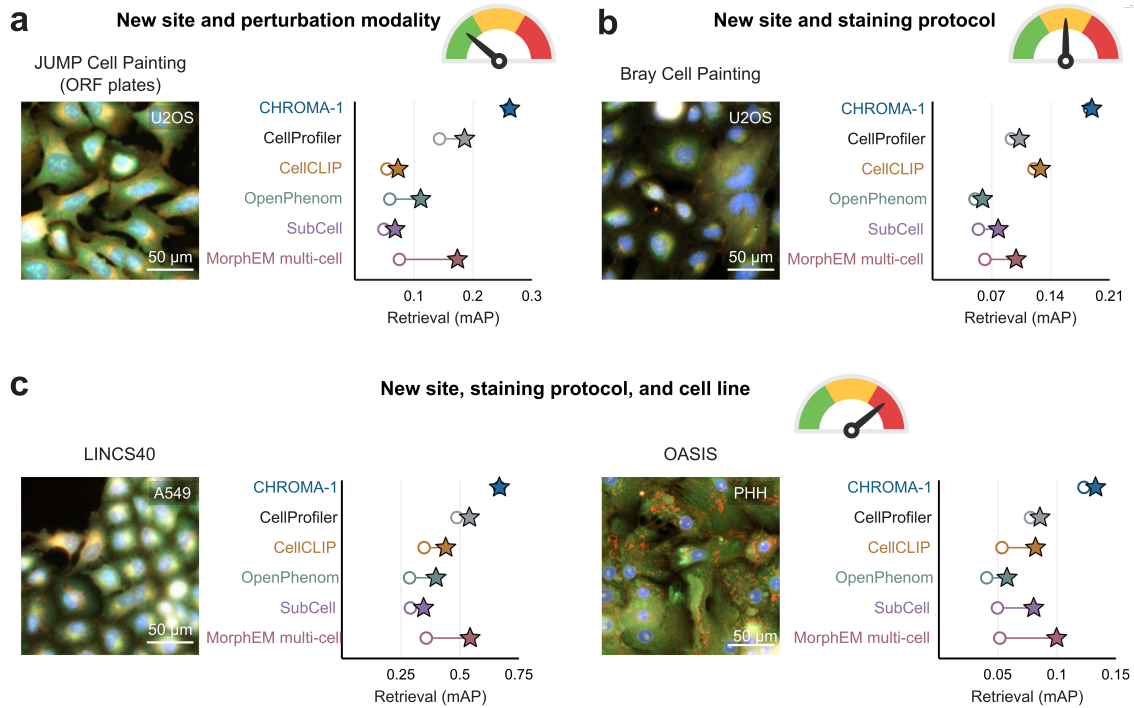


Figure 2: **Retrieval across four settings outside CHROMA-1’s training data.** (A) JUMP gene overexpression. (B) BBBC047. (C) LINCS A549 cells and OASIS primary human hepatocytes. Left: nonreplicate mAP without post hoc batch correction. Right: each comparator’s best score across tested batch corrections. Blue vertical lines mark CHROMA-1 without post hoc batch correction. Axes are shared within each row and vary across datasets. CellProfiler uses the feature preprocessing pipeline described by Arevalo et al. throughout. **BBBC047* was included in *CellCLIP*’s training. Representative control images show approximately $200 \times 200 \mu$ m fields with 50μ m scale bars.

3.1 Biological relationship retrieval transfers across training domains

This benchmark represents a substantial departure from CHROMA-1’s training conditions: HUVEC cells rather than U2OS, a different staining and acquisition protocol, and six imaging

channels rather than five [7, 28]. It also asks a different biological question from the preceding benchmarks. Instead of retrieving repeated instances of the same perturbation, profiles must connect distinct perturbations through known drug–target associations and gene–gene relationships. Following the published benchmark protocol [7], we measure drug–target retrieval using mean average precision and gene–gene retrieval using recall across CORUM [29], hu.MAP [30], Reactome [31] and STRING [32]. Figure 3A summarizes these complementary endpoints and Table 2 reports the individual endpoints.

CHROMA-1 exceeds every included comparator across all five endpoints, including OpenPhenom, Phenom-1 and Phenom-2. The training-data contrast makes this result particularly informative: RxRx3 [33] was used directly to train OpenPhenom and forms part of the broader microscopy collections underlying the Phenom models [19, 8, 9], while CHROMA-1 was trained entirely on a different dataset. CHROMA-1 surpasses these representations while transferring across both cellular and experimental conditions. This performance is achieved with a substantially smaller model and training budget, as illustrated in Figure 3B [8, 9].

These results show that CHROMA-1’s transfer capabilities extend beyond recognizing reproducible perturbation effects to recovering relationships between perturbations. Its consistent performance across drug–target associations, protein complexes and functional networks supports the use of its profiles to investigate shared biological mechanisms in unfamiliar experimental settings. The next question is whether this biological information also translates into stronger pharmacological predictions and fewer replicate wells needed to obtain them.

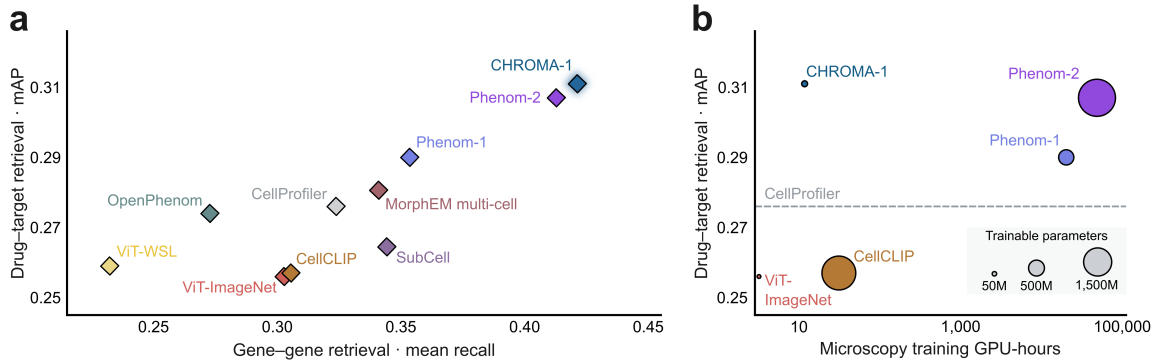


Figure 3: **Biological relationship retrieval and training resources on RxRx3.** (A) Drug–target mAP versus mean recall across CORUM, hu.MAP, Reactome and STRING. (B) Drug–target mAP versus microscopy training GPU-hours, with circle area proportional to encoder parameter count. Panel B includes models with supported aggregate training-resource figures; the dashed line marks CellProfiler. CHROMA-1 uses GH200 GPUs, Phenom-1 A100s and Phenom-2 H100s; CellCLIP combines A40 feature extraction with RTX 6000 training. ViT-ImageNet has no microscopy-training cost; upstream natural-image pretraining is excluded throughout. The horizontal axis is logarithmic above 10 GPU-hours and linear from zero to 10. The axis reports recorded GPU-hours. CellCLIP’s marker includes DINOv2-Giant, CrossChannelFormer and attention pooling [18]; Phenom resource figures follow the reported training budgets and encoder sizes [8, 9].

4 CHROMA-1 recovers pharmacological information with roughly half the replication

For screening applications, representation quality determines not only what can be predicted from Cell Painting images, but how much experimental replication is required to predict it. We

Table 2: **Biological relationship retrieval on RxRx3-core**. CHROMA-1 achieves the highest score across all five endpoints. OpenPhenom, Phenom-1 and Phenom-2 trained on microscopy collections containing RxRx3. Drug–target retrieval is measured by mAP; CORUM, hu.MAP, Reactome and STRING by recall. Higher is better in every column. Dashes indicate unavailable results. ViT-ImageNet provides a natural-image-pretrained baseline; ViT-WSL is a microscopy model trained with weak supervision, as described in the benchmark [7, 8].

Model	Drug–target mAP	CORUM Recall	hu.MAP Recall	Reactome Recall	STRING Recall
CHROMA-1	0.311	0.487	0.568	0.206	0.424
Phenom-2	0.307	0.486	0.553	0.197	0.415
Phenom-1	0.290	0.395	0.482	0.188	0.349
MorphEM multi-cell	0.281	0.405	0.461	0.149	0.348
CellProfiler	0.276	0.361	0.444	0.160	0.330
OpenPhenom	0.274	0.300	0.352	0.158	0.281
SubCell multi-cell	0.264	0.402	0.474	0.149	0.352
ViT-WSL	0.259	0.249	0.290	0.148	0.242
CellCLIP	0.257	0.354	0.416	0.145	0.307
ViT-ImageNet	0.256	0.342	0.420	0.144	0.305
CLOOME [34]	—	0.328	0.406	0.135	0.276
MolPhenix [35]	—	0.262	0.306	0.142	0.241

evaluate three complementary pharmacological endpoint families in held-out compounds. Protein activity comprises 24 experimentally measured target-and-mode assays [36, 37]; mechanism of action comprises 20 annotated mechanism classes [38, 39]; and safety pharmacology comprises 58 target-and-assay-mode endpoints from the Safety-77 panel [40, 41, 42]. Together, these test whether morphological profiles recover molecular activity, pharmacological mechanism, and safety-relevant target activity.

Across all three endpoint families, CHROMA-1 reaches CellProfiler’s predictive performance with roughly half the treatment wells at every evaluated budget (Figure 4A–C). This relationship is evaluated for 20 random selections of treatment wells for each budget. Thus, within a fixed treatment-well allocation, comparable predictive performance can be obtained with substantially less replication, leaving capacity for additional compounds, concentrations, or conditions. Alternatively, retaining the original replication budget yields stronger predictions: CHROMA-1 achieves higher macro AP at every matched budget across all three endpoint families (Figure 4D–F).

The replication advantage is consistent with stronger treatment signal in the profiles themselves. Compound–dose identity accounts for 43.6% of fitted CHROMA-1 profile variation, compared with 25.5% for CellProfiler after Harmony [17]. An independent variance model predicts that matching CellProfiler’s signal-to-noise ratio would require 0.44 CHROMA-1 wells per CellProfiler well, consistent with the roughly half-budget relationship observed in the pharmacological predictions. The variance model is descriptive and does not establish causal mediation, but provides independent evidence that CHROMA-1 concentrates more reproducible treatment-associated variation into each profile.

The gain is not explained by cell number. The protein-activity evaluation uses the benchmark of Seal et al. [36], designed for settings where cell count itself can be a powerful bioactivity baseline. Here, cell count alone is substantially less predictive than either full representation. CHROMA-1 retains its advantage when count features are appended, when their linear contribution is removed from the profiles, and among compounds whose cell counts remain within 15% of vehicle controls.

Interestingly, CHROMA-1’s advantage increases when conventional phenotypes are weak. Among

compounds in the weakest CellProfiler phenotype tertile, CHROMA-1’s advantage over CellProfiler is 7.75 AP points larger than among strong phenotypes for protein activity, and 6.25 points larger for safety pharmacology. To avoid circular reasoning, we define phenotype strength without using the activity labels and on disjoint wells.

Finally, CHROMA-1 is remarkably resilient to low negative-control allocations. With a single negative-control well per plate, its mean macro-AP loss across treatment budgets is only 0.1–0.4 percentage points for the three endpoint families. When both methods are restricted to two control wells per plate, CHROMA-1 remains essentially unchanged, while CellProfiler loses 1.3–2.5 macro-AP points relative to its full-control baseline. Thus, the treatment-well efficiency does not appear to require a large accompanying allocation of negative-control wells.

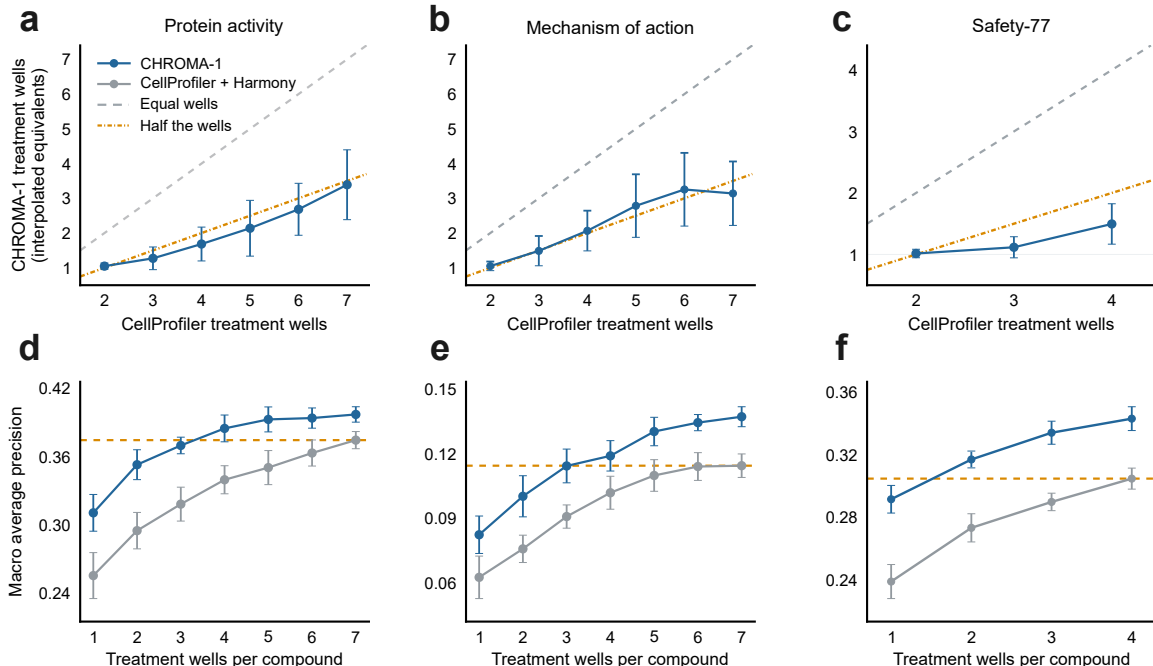


Figure 4: **Fewer treatment wells recover the same pharmacological information.** Columns show protein activity, mechanism of action and Safety-77. (A–C) CHROMA-1 well equivalents needed to reach CellProfiler at each reference budget, using linear interpolation between measured budgets within each of 20 matched well selections. Guide lines indicate equal and half as many wells. (D–F) CHROMA-1 achieves higher macro AP at every matched treatment-well budget. Shading indicates the gap between the mean curves; panel headings report the AP-point gain averaged equally across budgets. Points and bars in both rows show mean $\pm$ sample SD across the 20 selections. CHROMA-1 uses no post hoc batch correction; CellProfiler uses Harmony. Both retain their full control populations.

5 CHROMA-1 reveals mutagenicity signals missed by the CellProfiler baseline

Improved representations can expand the range of biological endpoints accessible from an imaging experiment, beyond improving performance on established tasks. Mutagenicity provides a demanding test of this possibility. The Ames assay measures the ability of a compound to induce mutations in bacteria, providing an independently measured safety endpoint that is biologically distinct from the morphological responses captured in human cells [43]. Predicting this activity from Cell Painting tests whether the images contain useful information about a compound’s mu-

tagenic potential. Prior work has predicted mutagenicity from Cell Painting [44], but selected positive and negative compounds using different morphological-activity thresholds, so prediction could largely reflect overall phenotype strength. We instead evaluate dose-qualified Ames outcomes without selecting compounds by morphological-response strength.

We linked compounds in BBBC047 to public Ames assay records from the National Toxicology Program’s Chemical Effects in Biological Systems database [45]. We evaluated whether a compound produces a positive Ames result at or below 10, 30 or 100 μg per plate. The lowest threshold focuses on compounds that induce mutations at low doses. Negative compounds were required to have been tested up to at least the corresponding dose. The three cohorts contain 87, 85 and 91 compounds, with 11, 12 and 19 positives, respectively. Using all available replicate wells, we compared CHROMA-1 without post hoc correction against CellProfiler with Harmony batch correction, evaluating predictions on held-out compounds. To make performance interpretable across the three endpoints, Figure 5 reports average precision relative to its expected value under random ranking. A value of one indicates random performance; higher values indicate enrichment of mutagenic compounds among the highest-ranked predictions.

CHROMA-1 shows a substantial predictive signal, while CellProfiler remains close to the random baseline. Across the three dose thresholds, CHROMA-1 achieves 2.89-, 2.33- and 1.75-fold enrichment over random, compared with 1.27-, 1.37- and 1.02-fold for CellProfiler. The separation is greatest at 10 μg per plate, where CHROMA-1 provides a nearly threefold enrichment for compounds active at low doses. CHROMA-1’s 95% confidence intervals remain above random at every threshold, while CellProfiler’s include the random baseline. CHROMA-1 improves average precision over CellProfiler by 128%, 70% and 72%, with paired comparisons supporting improvements across all three endpoints: the 95% confidence intervals for relative improvement are 23–243%, 2–172% and 21–151%, respectively.

These results extend the value of CHROMA-1 beyond reducing the experimental requirements for established pharmacological predictions. The same image collection supports an independently measured safety endpoint for which conventional morphological features provide limited predictive enrichment. This supports a broader role for improved morphological representations: opening existing image libraries to biological questions that were previously difficult to address, including the prioritization of compounds for mutagenicity testing.

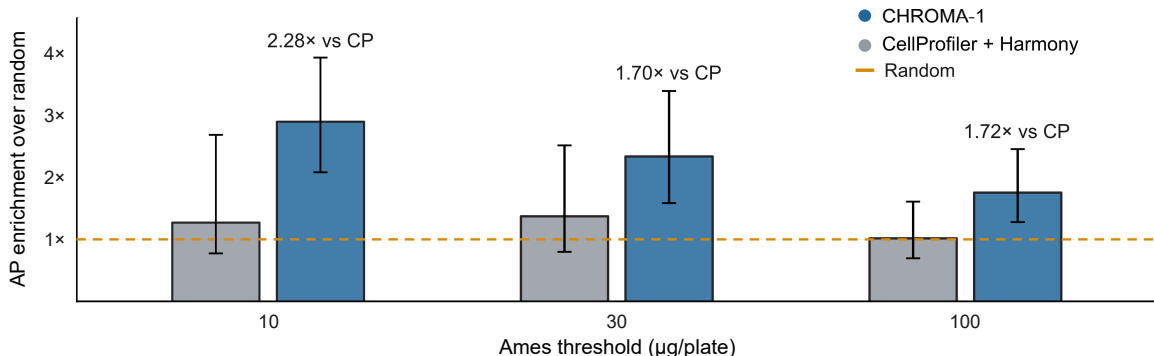


Figure 5: **Prediction of Ames mutagenicity from BBBC047 using all available replicate wells.** Bars show mean outer-fold average precision divided by the expected mean AP under random ranking, at Ames dose thresholds of 10, 30 and 100 μg per plate. CellProfiler uses the preprocessing described by Arevalo et al. followed by Harmony. The dashed line marks random performance. Error bars show pointwise 95% percentile intervals from 10,000 paired compound-bootstrap resamples, stratified by outer fold and class. Annotations above each pair report the ratio of CHROMA-1 AP to CellProfiler AP.

6 Discussion

CHROMA-1 expands the biological information accessible from morphological images while reducing the experimental resources needed to recover it. Its advantage spans cross-batch retrieval, transfer to unseen experimental settings and prediction of independently measured pharmacological endpoints. These evaluations establish complementary properties: preservation of biological relationships despite technical variation, generalization beyond the training domain and predictive utility for specific biological questions. Their agreement supports a broader conclusion than any individual benchmark: CHROMA-1 produces representations that remain useful as both the experimental context and the biological task change.

Our comparisons reaffirm recent benchmark findings that learned vision representations do not consistently improve on CellProfiler, particularly when applied outside their training conditions. Consistent with MorphoHELM and JUMP-lite [4, 5], CellProfiler remains a strong baseline across the evaluated settings, and even relatively modest experimental shifts expose weaknesses in competing learned representations. CellCLIP’s stronger performance on BBBC047, which was represented in its training data [18], further emphasizes the importance of distinguishing performance within a familiar domain from transfer. The central challenge is not simply to learn morphological features, but to demonstrate that those features retain their biological utility beyond the conditions in which they were learned.

CHROMA-1 addresses this challenge across several levels of experimental change. Without post hoc batch correction, it outperforms the tested alternatives on cross-batch biological retrieval even when each alternative receives its best-performing correction. Its advantage extends across changes in acquisition site, perturbation modality, staining protocol and cell system. On RxRx3-core, CHROMA-1 matches or surpasses the included state-of-the-art proprietary models, Phenom-1 and Phenom-2, across all five evaluated biological endpoints [7]. This comparison is particularly informative because RxRx3 belongs to the microscopy collections underlying the Phenom models [8, 9], whereas CHROMA-1 transfers from a different dataset, cell line and imaging protocol. CHROMA-1 achieves stronger biological retrieval despite a substantially greater shift from its training conditions.

The pharmacological evaluations establish the practical consequence of this representation quality. Across protein activity, mechanism of action and safety pharmacology, CHROMA-1 reaches CellProfiler’s macro performance with approximately 50% fewer replicate wells across the evaluated budgets. Experimental capacity could therefore be redirected toward additional compounds, concentrations or conditions. The Ames results demonstrate a distinct gain: CHROMA-1 extracts substantial predictive signal for an independently measured safety endpoint where CellProfiler remains close to random. These findings separate two benefits that are often conflated: recovering established biological information with fewer measurements, and recovering information that existing representations make difficult to access.

The smaller computational footprint makes these capabilities practical to develop and deploy. Relative to Phenom-1 and Phenom-2, CHROMA-1 uses approximately 1,700 and 4,100 times fewer recorded training GPU-hours, and 6 and 37 times fewer encoder parameters. At the measured inference throughput and nine sites per well, a 100,000-compound screen with one treatment well per compound corresponds to approximately 3.8 GPU-hours for treatment-image profiling.

CHROMA-1 makes each treatment well more informative. Its representations preserve biological relationships across experimental domains, support useful pharmacological predictions with roughly half the treatment replication, and reveal predictive signals that CellProfiler recovers poorly. For existing image libraries, this means more biological information from experiments already performed. For future screens, it offers a direct route to testing more compounds, con-

centrations, and conditions within the same treatment-well budget.

References

- [1] M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, S. M. Gustafsdottir, C. C. Gibson, and A. E. Carpenter, “Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes,” *Nature Protocols*, vol. 11, no. 9, pp. 1757–1774, 2016.
- [2] B. A. Cimini *et al.*, “Optimizing the Cell Painting assay for image-based profiling,” *Nature Protocols*, vol. 18, pp. 1981–2013, 2023. [Online]. Available: <https://doi.org/10.1038/s41596-023-00840-9>
- [3] D. R. Stirling *et al.*, “CellProfiler 4: improvements in speed, utility and usability,” *BMC Bioinformatics*, vol. 22, p. 433, 2021. [Online]. Available: <https://doi.org/10.1186/s12859-021-04344-9>
- [4] E. Hayir, L. Crawford, and A. X. Lu, “MorphoHELM: A Comprehensive Benchmark for Evaluating Representations for Microscopy-Based Morphology Assays,” Preprint, 2026. [Online]. Available: <https://arxiv.org/abs/2605.15383v1>
- [5] A. F. Muñoz *et al.*, “JUMP-lite: Compact, reproducible benchmarking of cell representations,” Preprint, 2026. [Online]. Available: <https://arxiv.org/abs/2608.07632v2>
- [6] J. Arevalo *et al.*, “Evaluating batch correction methods for image-based cell profiling,” *Nature Communications*, vol. 15, p. 6516, 2024. [Online]. Available: <https://doi.org/10.1038/s41467-024-50613-5>
- [7] O. Kraus, F. Comitani, J. Urbanik, K. Kenyon-Dean, L. Arumugam, S. Saberian, C. Wognum, S. Celik, and I. S. Haque, “RxRx3-core: Benchmarking drug-target interactions in High-Content Microscopy,” LMRL Workshop at ICLR, 2025. [Online]. Available: <https://openreview.net/pdf?id=0zQ0hIbuQd>
- [8] O. Kraus, K. Kenyon-Dean, S. Saberian, M. Fallah, P. McLean, J. Leung, V. Sharma, A. Khan, J. Balakrishnan, S. Celik *et al.*, “Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 757–11 768. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Kraus_Masked_Autoencoders_for_Microscopy_are_Scalable_Learners_of_Cellular_Biology_CVPR_2024_paper.html
- [9] K. Kenyon-Dean, Z. J. Wang, J. Urbanik, K. Donhauser, J. Hartford, S. Saberian, N. Sahin, I. Bendidi, S. Celik, J. S. R. Vera, M. Fay, I. S. Haque, and O. Kraus, “ViTally Consistent: Scaling Biological Representation Learning for Cell Microscopy,” in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, pp. 29 735–29 752. [Online]. Available: <https://proceedings.mlr.press/v267/kenyon-dean25a.html>
- [10] S. N. Chandrasekaran, J. Ackerman, E. Alix *et al.*, “JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations,” bioRxiv preprint, 2023.
- [11] E. Weisbart, A. Kumar, J. Arevalo *et al.*, “Cell painting gallery: an open resource for image-based profiling,” *Nature Methods*, vol. 21, pp. 1775–1777, 2024.

- [12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.
- [13] A. A. Kalinin, J. Arevalo *et al.*, “A versatile information retrieval framework for evaluating profile strength and similarity,” *Nature Communications*, vol. 16, p. 5181, 2025. [Online]. Available: <https://doi.org/10.1038/s41467-025-60306-2>
- [14] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- [15] M. D. Luecken *et al.*, “Benchmarking atlas-level data integration in single-cell genomics,” *Nature Methods*, vol. 19, pp. 41–50, 2022. [Online]. Available: <https://doi.org/10.1038/s41592-021-01336-8>
- [16] M. Büttner *et al.*, “A test metric for assessing single-cell RNA-seq batch correction,” *Nature Methods*, vol. 16, pp. 43–49, 2019. [Online]. Available: <https://doi.org/10.1038/s41592-018-0254-1>
- [17] I. Korsunsky *et al.*, “Fast, sensitive and accurate integration of single-cell data with Harmony,” *Nature Methods*, vol. 16, pp. 1289–1296, 2019. [Online]. Available: <https://doi.org/10.1038/s41592-019-0619-0>
- [18] M. Lu, E. Weinberger, C. Kim, and S.-I. Lee, “CellCLIP—Learning Perturbation Effects in Cell Painting via Text-Guided Contrastive Learning,” in *Advances in Neural Information Processing Systems*, vol. 38, 2025. [Online]. Available: https://papers.nips.cc/paper_files/paper/2025/hash/b4689a8e6ac04c0919f7162365f1248d-Abstract-Conference.html
- [19] Recursion, “OpenPhenom,” Released model and model card, evaluated checkpoint revision 0f92333685f6. [Online]. Available: <https://huggingface.co/recursionpharma/OpenPhenom>
- [20] A. Gupta *et al.*, “SubCell: Proteome-aware vision foundation models for microscopy capture single-cell biology,” Preprint, 2025. [Online]. Available: <https://doi.org/10.1101/2024.12.06.627299>
- [21] V. Agrawal *et al.*, “CHAMMI-75: Pre-training multi-channel models with heterogeneous microscopy images,” *ICLR*, 2026. [Online]. Available: <https://openreview.net/forum?id=SLjqdj3LPk>
- [22] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, vol. 177, pp. 1888–1902.e21, 2019. [Online]. Available: <https://doi.org/10.1016/j.cell.2019.05.031>
- [23] S. N. Chandrasekaran *et al.*, “Morphological map of under- and overexpression of genes in human cells,” *Nature Methods*, vol. 22, pp. 1742–1752, 2025. [Online]. Available: <https://doi.org/10.1038/s41592-025-02753-9>
- [24] M.-A. Bray *et al.*, “A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay,” *GigaScience*, vol. 6, p. giw014, 2017. [Online]. Available: <https://doi.org/10.1093/gigascience/giw014>
- [25] Broad Bioimage Benchmark Collection, “BBBC047, version 1: A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay.” [Online]. Available: <https://bbbc.broadinstitute.org/BBBC047>

- [26] G. P. Way *et al.*, “Morphology and gene expression profiling provide complementary information for mapping cell state,” *Cell Systems*, vol. 13, pp. 911–923.e9, 2022. [Online]. Available: <https://doi.org/10.1016/j.cels.2022.10.001>
- [27] J. D. Ewald *et al.*, “Cell Painting for cytotoxicity and mode-of-action analysis in primary human hepatocytes,” *Cell Systems*, vol. 17, p. 101566, 2026. [Online]. Available: <https://doi.org/10.1016/j.cels.2026.101566>
- [28] S. Celik, J.-C. Hütter, S. M. Carlos, N. H. Lazar, R. Mohan, C. Tillinghast, T. Biancalani, M. M. Fay, B. A. Earnshaw, and I. S. Haque, “Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data,” *PLOS Computational Biology*, vol. 20, no. 10, p. e1012463, 2024, supplementary Table S1. [Online]. Available: <https://doi.org/10.1371/journal.pcbi.1012463>
- [29] G. Tsitsiridis, R. Steinkamp, M. Giurgiu, B. Brauner, G. Fobo, G. Frishman, C. Montrone, and A. Ruepp, “CORUM: the comprehensive resource of mammalian protein complexes—2022,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D539–D545, 2023. [Online]. Available: <https://doi.org/10.1093/nar/gkac1015>
- [30] K. Drew, J. B. Wallingford, and E. M. Marcotte, “hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies,” *Molecular Systems Biology*, vol. 17, no. 5, p. e10016, 2021. [Online]. Available: <https://doi.org/10.15252/msb.202010016>
- [31] M. Milacic, D. Beavers, P. Conley *et al.*, “The Reactome Pathway Knowledgebase 2024,” *Nucleic Acids Research*, vol. 52, no. D1, pp. D672–D678, 2024. [Online]. Available: <https://doi.org/10.1093/nar/gkad1025>
- [32] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering, “The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D638–D646, 2023. [Online]. Available: <https://doi.org/10.1093/nar/gkac1000>
- [33] M. M. Fay, O. Kraus, M. Victors *et al.*, “RxRx3: Phenomics map of biology,” *bioRxiv* preprint, 2023.
- [34] A. Sanchez-Fernandez, E. Rumetshofer, S. Hochreiter, and G. Klambauer, “CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures,” *Nature Communications*, vol. 14, p. 7339, 2023.
- [35] P. Fradkin, P. Azadi, K. Suri, F. Wenkel, A. Bashashati, M. Sypetkowski, and D. Beaini, “How molecules impact cells: Unlocking contrastive PhenoMolecular retrieval,” in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
- [36] S. Seal, W. Dee, A. Shah *et al.*, “Counting cells can accurately predict small-molecule bioactivity benchmarks,” *Nature Communications*, vol. 17, p. 2436, 2026. [Online]. Available: <https://www.nature.com/articles/s41467-026-68725-5>
- [37] Eve Bio, “Pharmome data, release 4,” 2025. [Online]. Available: <https://data.evebio.org/>
- [38] Broad Bioimage Benchmark Collection, “BBBC036, version 1: Human U2OS cells—profiling bioactive compounds using Cell Painting.” [Online]. Available: <https://bbbc.broadinstitute.org/BBBC036>

- [39] S. M. Corsello, J. A. Bittker, Z. Liu *et al.*, “The Drug Repurposing Hub: a next-generation drug library and information resource,” *Nature Medicine*, vol. 23, pp. 405–408, 2017. [Online]. Available: <https://www.nature.com/articles/nm.4306>
- [40] R. J. Brennan, S. Jenkinson, A. Brown *et al.*, “The state of the art in secondary pharmacology and its impact on the safety of new medicines,” *Nature Reviews Drug Discovery*, vol. 23, pp. 525–545, 2024. [Online]. Available: <https://www.nature.com/articles/s41573-024-00942-3>
- [41] J. J. Sutherland, D. Yonchev, A. Fekete, and L. Urban, “A preclinical secondary pharmacology resource illuminates target-adverse drug reaction associations of marketed drugs,” *Nature Communications*, vol. 14, p. 4323, 2023. [Online]. Available: <https://www.nature.com/articles/s41467-023-40064-9>
- [42] F. Schmidt, R. J. Brennan, S. Jenkinson, and J.-P. Valentin, “Shaping secondary pharmacology panels of the future: evolving target selection criteria for safety panels,” *Nature Reviews Drug Discovery*, vol. 24, pp. 482–484, 2025. [Online]. Available: <https://www.nature.com/articles/s41573-025-01184-7>
- [43] OECD, *Test No. 471: Bacterial Reverse Mutation Test*, ser. OECD Guidelines for the Testing of Chemicals, Section 4. Paris: OECD Publishing, 2020. [Online]. Available: <https://doi.org/10.1787/9789264071247-en>
- [44] N. Cerisier, E. Truong, T. Watanabe, T. Oshiro, T. Takahashi, S. Ito, and O. Taboureau, “Computational prediction of mutagenicity through comprehensive cell painting analysis,” *Mutagenesis*, vol. 40, no. 4, pp. 560–574, 2025. [Online]. Available: <https://academic.oup.com/mutage/article/40/4/560/8221735>
- [45] National Institute of Environmental Health Sciences, “Chemical Effects in Biological Systems: Bacterial Reverse Mutation Test (Ames Test) Individual Test Data,” bacterial reverse mutation individual-test export accessed September 20, 2026. [Online]. Available: <https://cebs-ext.niehs.nih.gov/datasets/search/ames>