MorphGen: Controllable and Morphologically Plausible Generative Cell-Imaging

Berker Demirel [1 2] **Marco Fumero** [1] **Theofanis Karaletsos** [2] **Francesco Locatello** [1 2]

Abstract

Simulating in silico cellular responses to interventions is a promising direction to accelerate high-content image-based assays, critical for advancing drug discovery and gene editing. To support this, we introduce MorphGen, a state-of-the-art diffusion-based generative model for fluorescent microscopy that enables controllable generation across multiple cell types and perturbations. To capture biologically meaningful patterns consistent with known cellular morphologies, MorphGen is trained with an alignment loss to match its representations to the phenotypic embeddings of OpenPhenom, a biological foundation model. Unlike prior approaches that compress multichannel stains into RGB images –thus sacrificing organelle-specific detail– MorphGen generates the complete set of fluorescent channels jointly, preserving per-organelle structures. Despite modeling four cell types and generating the complete set of fluorescent channels, MorphGen achieves an FID score over 35% lower than the prior state-of-the-art MorphoDiff, which only generates RGB images for a single cell type.

1. Introduction

Deep generative models are emerging tools for simulating cellular behavior in computational biology, with early works in modeling gene expression profiles (Bereket & Karaletsos, 2024) and more recently synthesizing microscopy images (Navidi et al., 2025; Palma et al., 2025), which can be easily collected at scale. These models offer the potential to create *in silico* surrogates of biological experiments. This is a critical step in the vision of *Virtual Cells* (Bunne et al., 2024): a generative instrument capable of populating diverse cellular contexts and emulating the effects of genetic or chemical interventions. Realizing such a system could accelerate biological discovery by producing high-quality hypotheses without the time and cost constraints of exhaustive wet-lab experiments. As a practical step toward this vision, we focus on phenotypic image generation under experimentally defined perturbations.

However, current image generators fall short of these goals: (i) they operate at low resolution and rely on outdated architectures (Palma et al., 2025); (ii) they collapse six-channel fluorescence stacks into lossy RGB; and (iii) they are restricted to a single cell type and modest-sized datasets (Navidi et al., 2025). As a result, they miss out on both fine-grained morphological analysis and realism. Instead, we posit that a generative model should maintain local biological information, at the individual fluorescence level. Further, restricting generation to a single cell type limits the model's generality, posing limitations toward applications. For a more detailed related work, please refer to Appendix A.

We present **MorphGen**, a generative model that addresses these gaps and supports generation across many perturbations and four cell lines. Its contributions are:

- **Organelle-level generation.** MorphGen synthesizes native fluorescence channels directly, avoiding RGB conversion and preserving subcellular detail.
- **Controllable synthesis.** A structured latent space disentangles perturbation and cell-type factors, enabling compositional and targeted generation.
- **Scalable training.** MorphGen is trained at full resolution on the RxRx1 dataset (Sypetkowski et al., 2023), covering four cell types and 125K images at 512×512.
- **Biological fidelity.** An alignment loss guided by OpenPhenom (Kraus et al., 2024) embeddings ensures biologically meaningful features, supported by downstream validation using CATEs.

Figure 1 illustrates the visual fidelity of our model across four representative cell-type/perturbation pairs. To the best of our knowledge, MorphGen is the first generator that delivers high-resolution, organelle-aware, and biologically faithful Cell Painting images at scale.

[1]Institute of Science and Technology, Klosterneuburg, Austria [2]Chan Zuckerberg Initiative, California, USA. Correspondence to: Berker Demirel <berker.demirel@ist.ac.at>.

SIM 2025 Workshop at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada.

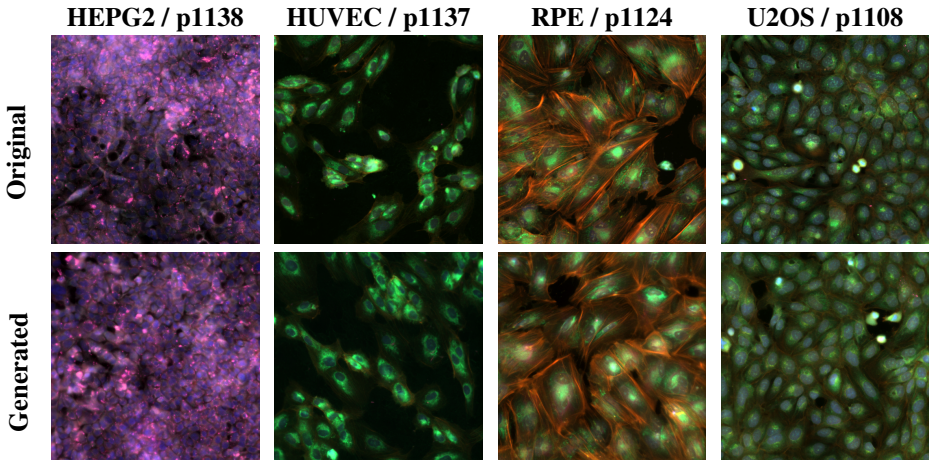


Figure 1: Original (top row) and generated (bottom row) images for various cell-type/perturbation pairs from the RxRx1 dataset. MorphGen generates high fidelity images across different cell-types and perturbations.

2. MorphGen

MorphGen is a generative model for synthesizing high-resolution, biologically meaningful cell images across diverse perturbations and cell types. It combines a pretrained VAE with a latent diffusion model tailored for multi-channel Cell Painting data. To address challenges of high-dimensional latents (e.g., from channel concatenations) and enhance biological fidelity, we use a REPA-inspired (Yu et al., 2025) alignment loss and train the diffusion model using features from a biological foundation model.

We consider a conditional latent diffusion setting for high-resolution, multi-channel fluorescence microscopy images. Let $\mathcal{X} \subset \mathbb{R}^{6\times H\times W}$ denote the image space of six-channel Cell Painting images.

Organelle-aware processing. Since the pretrained VAE encoder is designed for three-channel RGB images, we adapt each grayscale input channel by stacking it three times along the channel dimension. Let $\mathbf{x}^{(c)} \in \mathbb{R}^{1\times H\times W}$ be the c-th channel of an image $\mathbf{x} \in \mathcal{X}$. We define its RGB-stacked version as $\tilde{\mathbf{x}}^{(c)} \in \mathbb{R}^{3\times H\times W}$. This design allows us to encode organelle-specific fluorescence channels independently, preserving its biological specificity.

Each stacked channel $\tilde{\mathbf{x}}^{(c)}$ is passed through a frozen pretrained VAE encoder E_{VAE} to obtain a compressed latent representation: $\mathbf{z}^{(c)} = E_{\text{VAE}}(\tilde{\mathbf{x}}^{(c)}) \in \mathbb{R}^{4\times H'\times W'}$, where H' and W' denote the spatial resolution of the VAE latent space. We then concatenate the six channel-wise latents along the channel dimension to form the full latent representation: $\mathbf{z} = \text{concat}(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(6)}) \in \mathbb{R}^{24\times H'\times W'}$.

Joint diffusion process. The concatenated latent $\mathbf{z}$ serves as the input to a latent diffusion model parameterized by a Scalable Interpolant Transformer (SiT) (Ma et al., 2024). Conditioning is achieved through the combination of perturbation, cell type, and diffusion timestep embeddings. Let $p \in \mathcal{P}$ and $ct \in \mathcal{CT}$ denote the perturbation and cell type labels, respectively, which are mapped to learnable embeddings $\mathbf{e}_p, \mathbf{e}_{ct} \in \mathbb{R}^d$. With the timestep embedding $\mathbf{e}_t$, the conditioning vector $\mathbf{c} = \mathbf{e}_p + \mathbf{e}_{ct} + \mathbf{e}_t$ is used as the cross-attention context in SiT.

Following Karras et al. (2022a), the forward diffusion process generates noisy latent samples by interpolating clean latents $\mathbf{z}_0$ with Gaussian noise. At a random timestep $t \in [0, T]$, this interpolation is given by $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, where α_t and σ_t are deterministic scaling factors with boundary conditions $\alpha_0 = \sigma_T = 1$ and $\alpha_T = \sigma_0 = 0$. The diffusion model predicts the velocity $\mathbf{v}_t$ of the trajectory, defined as the time derivative of the latent:

$$\mathbf{v}_t = \frac{d\mathbf{z}_t}{dt} = \dot{\alpha}_t \mathbf{z}_0 + \dot{\sigma}_t \boldsymbol{\epsilon}.$$

Given the noisy latent $\mathbf{z}_t$ and conditioning vector $\mathbf{c}$, the Scalable Interpolant Transformer (SiT) f_θ estimates this velocity: $\hat{\mathbf{v}}_t = f_\theta(\mathbf{z}_t, \mathbf{c})$, and is trained via mean squared error against the ground-truth velocity:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0, t, \boldsymbol{\epsilon}} \left[\| f_\theta(\mathbf{z}_t, \mathbf{c}) - (\dot{\alpha}_t \mathbf{z}_0 + \dot{\sigma}_t \boldsymbol{\epsilon}) \|_2^2 \right].$$

Incorporating biological representations. To improve the biological fidelity, we use REPA with OpenPhenom (Kraus et al., 2024) features during training. The alignment loss guides the model toward biologically meaningful representations and therefore, improve the image quality. Given a clean image $\mathbf{x}$, we extract reference patch-level embeddings: $y^\star = F(\mathbf{x}) \in \mathbb{R}^{N\times d'}$, where N is the number of patches and d' is the embedding dimension. Let $h_t^{(k)} \in \mathbb{R}^{N\times d}$ denote the representations at layer k of the SiT at timestep t. This hidden representation is projected through a learnable

MLP h_ϕ into dimension d' to align with $y^\star$. The REPA loss encourages alignment via cosine similarity:

$$\mathcal{L}_{\text{REPA}} = -\frac{1}{N}\sum_{n=1}^{N} \text{sim}\left(y_n^\star, h_\phi(h_{t,n}^{(k)})\right).$$

The total loss is $\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda\mathcal{L}_{\text{REPA}}$, where λ balances the alignment loss.

Sampling process. At inference, we use the Euler–Maruyama sampler to generate images from Gaussian noise in the concatenated latent space. At each noise level, the trained SiT predicts the drift that guides the latent toward a clean signal. Once the final step is reached, we split the $24 \times H' \times W'$ tensor back into six channel-specific latent representations. Each of these is decoded separately through the VAE decoder, yielding six RGB stacks of size $3\times H\times W$. We then collapse each stack to single grayscale channel by averaging its three color planes, and recombine all six to form the final $6 \times H \times W$. This inverse procedure enables six-channel fluorescent Cell Painting image generation.

3. Experiments

We evaluate MorphGen's quality through comparisons with prior models, perturbation- and cell-type conditioned generation, organelle-specific synthesis, and analysis using OpenPhenom features (Kraus et al., 2024). Dataset details, evaluation setup, and ablations are in Appendix B.1,B.2 and B.3.

3.1. Comparison with state-of-the-art

Table 1: FID and KID scores (lower is better) on HUVEC cell type. MorphGen significantly outperforms models that are trained to only generate HUVEC cells.

Method	FID ↓	KID ↓
Stable Diffusion	115	0.11
MorphoDiff	78	0.05
MorphGen (Ours)	**50.2**	**0.01**

Results. Table 1 demonstrates that, even under MorphoDiff's constrained HUVEC-only, RGB-mapped evaluation, MorphGen achieves substantial improvements: it reduces FID by 64.8 and 27.8, and KID by 0.10 and 0.04, compared to Stable Diffusion and MorphoDiff, respectively. Such large gains under a constrained setting highlight MorphGen's fidelity. Moreover, Figure 1 provides complementary qualitative evidence: across multiple cell-type/perturbation pairs, MorphGen's outputs faithfully reproduce the original morphology and texture, visually corroborating our quantitative results.

Table 2: FID and KID scores across cell types. MorphGen enables channel-wise generation for all four cell types

Metric	Nucleus	ER	Actin	Cyto	Nucleolus	Mito
FID ↓	27.6	48.1	57.6	49.6	43.6	59.0
KID ↓	0.010	0.011	0.015	0.013	0.012	0.012

3.2. Additional Capabilities

Organelle-Specific Generation We repeat the same procedure across all four cell types, generating images while matching the original cell-type distribution. As shown in Table 2, the nucleus channel yields the lowest FID (27.6) and KID (0.010), making it the easiest to model. In contrast, actin (FID 57.6, KID 0.015) and mitochondria (FID 59, KID 0.012) are comparatively harder. Importantly, even under this more general, multi-cell-type evaluation, MorphGen not only outperforms MorphoDiff's RGB-only HUVEC baseline, but also beats it on every individual channel. We demonstrate the qualitative comparison in Appendix D.

Table 3: Cell-type specific results. MorphGen is capable of generating high-fidelity images for different cell types.

Cell Type	FID ↓	KID ↓
HEPG2	39.4	0.016
HUVEC	29.8	0.006
RPE	33.6	0.007
U2OS	37.7	0.017

Cell-Type-Specific Generation. To assess MorphGen under more natural data distributions, we randomly sample images by cell type without conditioning on perturbations—avoiding any data augmentation to reach a fixed sample count. Table 3 reports FID and KID on the resulting RGB-converted images. MorphGen achieves its best scores on HUVEC and maintains strong performance across the other cell types. These cell-type-specific results outperform our earlier experiments, which required aggressive augmentation (random flips and rotations) to inflate small-perturbation sets to ensure compatibility with MorphoDiff, at the cost of introducing bias into the real data distribution. By contrast, when following the natural data distribution without artificial augmentation, our model's performance excels further, demonstrating MorphGen's superior fidelity.

3.3. Morphology Analysis with OpenPhenom Features

Feature Extraction and PCA Visualization. To qualitatively assess the biological plausibility of our generated images, we extract features using the foundation model OpenPhenom. We focus on the four most frequent perturbations in the dataset—including the control (p1138)—and visualize both real and generated samples in a shared PCA embedding space. Figure 2. shows that (i) real and

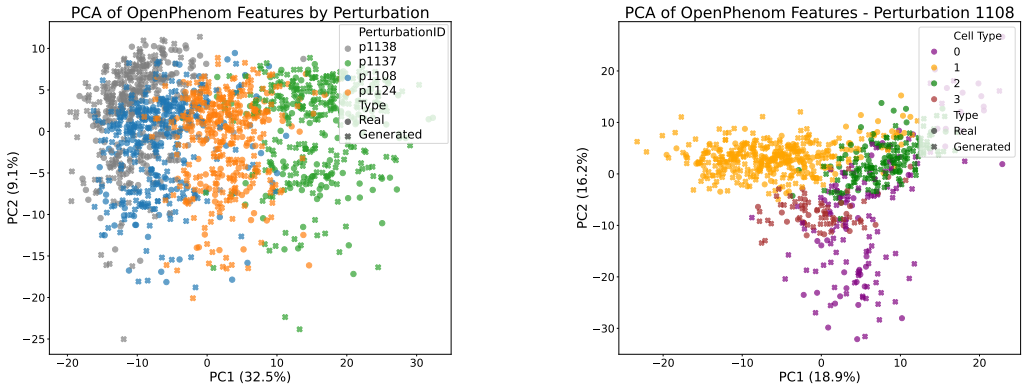


Figure 2: PCA of OpenPhenom features from real and generated images. Left: frequent perturbations (incl. control p1138) for HUVEC, colored by perturbation. Right: perturbation 1108 across cell types. Marker shapes denote real vs. generated.

generated embeddings largely overlap within the same perturbation class—indicating that generated images reproduce morphology faithfully, and (ii) perturbation-specific deviations from the control are similarly captured in both real and generated distributions—evident from the clear color separation. These patterns suggest that our generative model successfully encodes biologically relevant phenotypic variation while preserving class-level consistency.

CATE: Average Treatment Effect in Feature Space. To quantitatively validate that our generated images capture biologically meaningful perturbation effects (Bereket & Karaletsos, 2024) at the population level, we compute the Conditional Average Treatment Effect (CATE) between control and perturbed samples using OpenPhenom features. We use OpenPhenom features to represent cellular morphology and denote the image-level embedding as Y, obtained by averaging patch-level representations across the image. Given a perturbation p, we define CATE as associational difference between a treated population and a control group for a specific cell type:

$$\text{CATE}(p) = \Big\| \mathbb{E}[Y \mid P = \text{control},\ ct = \text{HUVEC}] - \mathbb{E}[Y \mid P = p,\ ct = \text{HUVEC}] \Big\|^2$$

This metric captures the squared Euclidean distance between the average feature vectors of the control group (p1138) and a perturbed group p. We compute the CATE separately for real and generated samples across the three most common perturbations: 1108, 1124, and 1137. While clearly the image-level embeddings do not correspond directly to biological quantities where the treatment effect can immediately be interpreted, Kraus et al. (2024) showed that the OpenPhenom features are very strong predictors of the CellProfiler (Carpenter et al., 2006) features. Therefore, we estimate the Average Treatment Effect in feature space, as the associational difference will carry over to any downstream morphological predictor (Cadei et al., 2024; 2025).

Table 4: Conditional Average Treatment Effect (CATE) between control (p1138) and perturbed samples, computed using real and generated HUVEC images.

Comparison	$\text{CATE}_{\text{real}}$	CATE_{gen}	$\Delta\text{CATE}\downarrow$
p1138 vs p1137	7.85	7.41	0.43
p1138 vs p1124	2.13	2.31	0.18
p1138 vs p1108	0.44	0.38	0.06

As shown in Table 4, p1137 results in the largest morphological deviation from the control, while p1108 has the smallest effect. These magnitudes align well with the spatial patterns in Figure 2. The close agreement between CATE values computed from real and generated images further indicates that our model reliably captures biologically meaningful perturbation effects. While alignment with OpenPhenom features may be expected due to the alignment loss, this loss operates in the latent space of a frozen VAE, so preservation of morphological features is not guaranteed. Our results confirm that they are, and that population-level statistical associations closely match those from real images.

4. Conclusion

We introduce MorphGen, a generative model that synthesizes high-resolution, six-channel Cell Painting images while preserving biologically meaningful structure across diverse perturbations and cell types. Trained at scale on the full RxRx1 dataset, MorphGen leverages a novel alignment loss guided by embeddings from a microscopy-specific foundation model. Beyond visual fidelity, features from generated images capture population trends, with CATEs closely matching those from real data—supporting their use in downstream phenotypic analysis. As future work, we aim to enable instance-based conditioning by learning embeddings from single exemplars, even in unseen settings. Despite limitations, MorphGen marks a step toward virtual instruments that accelerate hypothesis generation and experimental design in functional genomics and drug discovery.

Acknowledgements

This work was supported by the Chan Zuckerberg Initiative (CZI) through the External Residency Program. We thank CZI for the opportunity to participate in this program, which enabled close collaboration and access to critical resources. We also acknowledge the CZI AI Infrastructure Team for their support with the GPU cluster used to train our models.

References

AI, S. sd-vae-ft-mse. https://huggingface.co/stabilityai/sd-vae-ft-mse, 2022. Fine-tuned VAE decoder for Stable Diffusion, optimized with MSE + LPIPS loss.

Bereket, M. and Karaletsos, T. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *Advances in Neural Information Processing Systems*, 36, 2024.

Biewald, L. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D. B., et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Cadei, R., Lindorfer, L., Cremer, S., Schmid, C., and Locatello, F. Smoke and mirrors in causal downstream tasks. *Advances in Neural Information Processing Systems*, 37: 26082–26112, 2024.

Cadei, R., Demirel, I., De Bartolomeis, P., Lindorfer, L., Cremer, S., Schmid, C., and Locatello, F. Causal lifting of neural representations: Zero-shot generalization for causal inferences. *arXiv preprint arXiv:2502.06343*, 2025.

Caie, P. D., Walls, R. E., Ingleston-Orme, A., Daya, S., Houslay, T., Eagle, R., Roberts, M. E., and Carragher, N. O. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular cancer therapeutics*, 9(6):1913–1926, 2010.

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7:1–11, 2006.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gugger, S., Debut, L., Wolf, T., Schmid, P., Mueller, Z., Mangrulkar, S., Sun, M., and Bossan, B. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022a.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022b.

Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., and Laine, S. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv e-prints*, pp. arXiv–1312, 2013.

Kraus, O., Kenyon-Dean, K., Saberian, S., Fallah, M., McLean, P., Leung, J., Sharma, V., Khan, A., Balakrishnan, J., Celik, S., et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11757–11768, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., and Xie, S. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Navidi, Z., Ma, J., Miglietta, E., Liu, L., Carpenter, A. E., Cimini, B. A., Haibe-Kains, B., and WANG, B. Morphodiff: Cellular morphology painting with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PstM8YfhvI.

Palma, A., Theis, F. J., and Lotfollahi, M. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16(1):505, 2025.

Phillips, L. Cellrep: A multichannel image representation learning model. In *1st CVPR Workshop on Computer Vision For Drug Discovery (CVDD): Where are we and What is Beyond?*, 2025.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sypetkowski, M., Rezanejad, M., Saberian, S., Kraus, O., Urbanik, J., Taylor, J., Mabey, B., Victors, M., Yosinski, J., Sereshkeh, A. R., et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4285–4294, 2023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., and Xie, S. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

A. Related Work on Generative Models in Microscopy-Based HCS

In this section we discuss the recent work that attempts controllable generation of Cell-Painting images – MorphoDiff (Navidi et al., 2025) and IMPA (Palma et al., 2025). Both aim to illustrate the morphological response of a given perturbation, but they differ in model class, channel handling, resolution and biological evaluation. Here, we outline the essentials and provide a comparison with key design choices in MorphGen.

Morphodiff. MorphoDiff adapts a Stable-Diffusion (Rombach et al., 2022) latent DDPM (Ho et al., 2020) to Cell Painting. Perturbations are encoded with scGPT embeddings (Cui et al., 2024). Six fluorescence channels are projected into RGB through an irreversible compression that merges organelle-specific cues (Phillips, 2025) to remain compatible with the pretrained Stable Diffusion VAE (Kingma & Welling, 2013). The model trains on full resolution images from a single cell type in RxRx1 (Sypetkowski et al., 2023) and since unannotated perturbations lack scGPT indices the authors discard those images, limiting its general applicability, thus the model explores only a single factor of variation –the annotated perturbations.

IMPA. IMPA treats perturbation or batch as "style" and performs image-to-image translation with an AdaIN-conditional GAN (Goodfellow et al., 2014; Huang & Belongie, 2017). To fit the GAN, native 512×512 Cell-Painting images are cropped to 96×96 patches, sharply reducing spatial resolution and blurring organelle-level signals. The study trains only on the U2OS cell line from RxRx1 for batch-effect removal and small BBBC/JUMP-CP subsets (Caie et al., 2010) for perturbations. Therefore, cross-cell-type generation is untested. Compared to diffusion models, the GAN backbone offers lower fidelity (Karras et al., 2022b) and less stable training (Lucic et al., 2018), making IMPA less suited to high-resolution virtual phenotyping.

Comparison to MorphGen. Both prior methods leave critical gaps that MorphGen closes. Unlike MorphoDiff's irreversible RGB compression and IMPA's 96×96 down-sampling, MorphGen keeps every fluorescence channel intact by wrapping each grayscale slice in a three-channel latent and running diffusion jointly across all six channels. The latents are then split and decoded per channel, preserving organelle detail at the native 512×512 scale. The resulting higher latent dimensionality is tamed with a representation alignment loss –adapted from REPA (Yu et al., 2025)– but driven by OpenPhenom embeddings (Kraus et al., 2024), instead of generic vision features. This alignment loss stabilizes training and sharpens biological fidelity. Because our perturbation and cell-type embeddings are learned directly from images, MorphGen uses *all* RxRx1 plates (four cell lines, all perturbations), whereas MorphoDiff discards unlabelled perturbations while working only on HUVEC and IMPA is limited to U2OS. Together, full-channel diffusion, alignment loss and data-driven conditioning give MorphGen higher resolution, multi-cell-type generality and tighter biological concordance than either earlier model.

B. Experimental Details

B.1. Dataset

RxRx1 dataset. This dataset is a large-scale, high-resolution collection of fluorescence microscopy images designed to support the study of phenotypic cellular responses to gene knockdowns and to benchmark batch effect correction methods (Sypetkowski et al., 2023). It comprises $125,510$ images from four human cell types (HUVEC, RPE, HepG2 and U2OS), each exposed to one of $1,108$ siRNA treatments targeting distinct genes, along with 30 non-targeting control conditions. Imaging is performed using a modified Cell Painting assay, generating six-channel 512×512 pixel images that visualize major subcellular structures including the nucleus, endoplasmic reticulum, actin cytoskeleton, nucleoli, mitochondria and golgi apparatus. By capturing morphological changes induced by gene-specific knockdowns, RxRx1 serves as a challenging benchmark for models aiming to generalize across perturbations, cell types and experimental batches.

B.2. Evaluation Setup

Metrics. We report Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). Unlike (Navidi et al., 2025), we do not report the FID and KID scores divided by a factor of 100, but rather report the unscaled value as typical in computer vision (Karras et al., 2022a; Ma et al., 2024). Every score is computed from 500 generated versus 500 real images. All experiments are repeated with three random seeds; we report the mean.

- **Perturbation-level** (Sec. 3.1): To ensure a fair comparison with MorphoDiff, we adopt their RxRx1 protocol. Metrics are independently computed for randomly selected 50 siRNAs in HUVEC; then averaged across perturbations. Although MorphGen natively generates full 6-channel images across multiple cell lines, we restrict it to HUVEC-only genera-

tion—matching MorphoDiff's capacity—and convert our outputs to the RGB space using Recursion's visualization script (Sypetkowski et al., 2023).

- **Organelle-specific** (Sec. 3.2): metrics are computed using the same 50 siRNAs but evaluated in each of the four cell types and in every single channel representing organelles
- **Cell-type-level** (Sec. 3.1): metrics are computed per cell type without perturbation conditioning

Augmentation policy. When the real dataset for a given perturbation contained < 500 examples, we followed (Navidi et al., 2025) and synthetically expanded it using random flips and 90° rotations.

B.3. Implementation Details

MorphGen is trained as a latent-diffusion model operating on SD-VAE (Rombach et al., 2022) latents of six-channel RxRx1 (Sypetkowski et al., 2023) training set images of 512×512 resolution. Each single-channel grayscale image is stacked into a 3-channel RGB format, scaled to $[-1, 1]$, encoded with the public stabilityai/sd-vae-ft-mse VAE (AI, 2022), and rescaled by the SD constants $(0.18215, 0)$. The diffusion backbone is a Scalable Interpolant Transformer (SiT XL/2) (Ma et al., 2024). During training, OpenPhenom (Kraus et al., 2024) embeddings of the raw image are injected via a REPA-style projection loss (weight = 0.5), mirroring the original REPA formulation (Yu et al., 2025).

Optimization uses AdamW (β = 0.9/0.999, lr = 1×10^{-4}) (Loshchilov & Hutter, 2017) in mixed precision (fp16 or bf16) under HuggingFace Accelerate (Gugger et al., 2022); TF32 kernels can be enabled for additional speed. An EMA (Karras et al., 2024) shadow network (decay = 0.9999) is maintained for sampling. Sampling uses a 50-step Euler–Maruyama schedule (Ma et al., 2024).

Training runs with a batch size of 16 for up to 400 k steps using 8 H100 GPUs. Checkpoints are written every 50 k steps, and the best model is chosen by the average FID across distributed workers, computed on 100 real vs. generated images of pre-selected perturbations using Inception-V3 (Szegedy et al., 2016) features. Logging and image grids are tracked in Weights and Biases (Biewald, 2020), and all hyper-parameters, metrics, and checkpoints are stored for full reproducibility.

C. Ablation Study

Morphgen without the alignment loss. Table 5 presents an ablation study evaluating the effect of incorporating an alignment loss on OpenPhenom features. While Yu et al. (2025) introduced a similar alignment strategy to accelerate diffusion training, we instead leverage it to guide the model toward learning biologically meaningful representations during generation. With the alignment loss, MorphGen achieves more than a 10% reduction in FID and over a 20% reduction in KID, indicating a substantial improvement in image fidelity. These results support the effectiveness of our proposed approach in integrating biological priors through pretrained features during training.

Table 5: Ablation study on the alignment loss. We compare models trained without (top row) and with (bottom row) alignment regularization. FID and KID scores (lower is better) are reported for 50 randomly sampled perturbations from the HUVEC cell type.

Method	FID↓	KID↓
MorphGen wo/ align.	56.87 ± 3.35	0.023 ± 0.001
MorphGen (Ours)	$\mathbf{50.2 \pm 2.45}$	$\mathbf{0.018 \pm 0.000}$

Table 6: Comparison of reconstruction fidelity across different channel processing strategies using a frozen VAE.

Channel Processing	MSE↓
RGB	7.13×10^{-4}
Organelle-aware	4.93×10^{-5}
Organelle-aware + RGB	4.06×10^{-4}

Channel-wise Processing. To evaluate how different preprocessing strategies affect reconstruction fidelity using a pretrained VAE, we compared three settings. In the baseline setup (RGB), all six fluorescence channels are first compressed into an RGB image before encoding, as done in prior work like MorphoDiff. In contrast, our organelle-aware strategy (MorphGen) processes each channel independently by replicating it across RGB channels to match the VAE's input

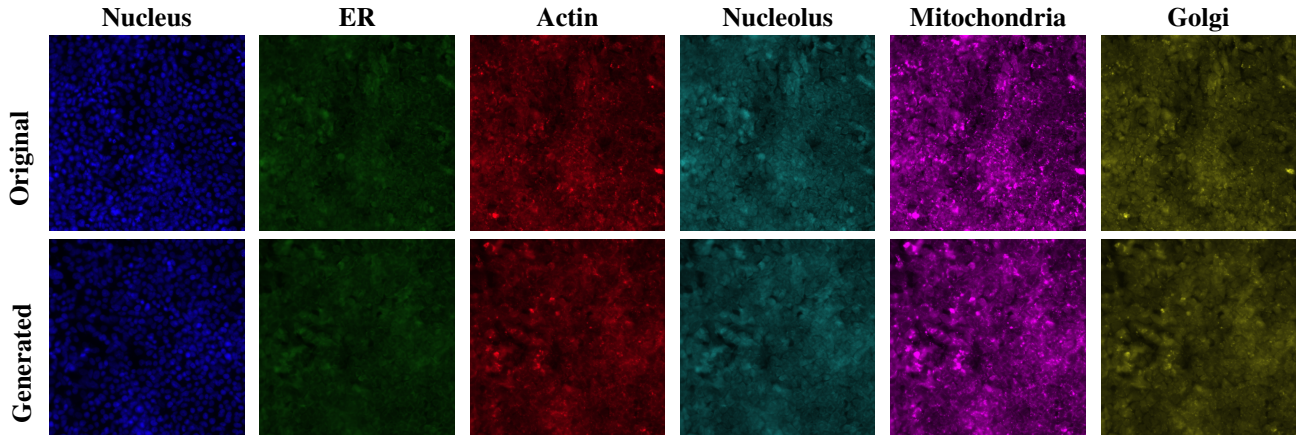


Figure 3: Comparison of original and generated fluorescence images for each organelle in a control HEPG2 cell. Our model reconstructs the six distinct fluorescent channels using RxRx1-recommended colormaps, preserving morphology across subcellular structures. Image best seen in color. Generated images are not cherry-picked, and we selected original images that are neighbors of the generated ones for visualization.

expectations. We then consider two ways of reporting reconstruction loss. First, we compute the per-channel reconstruction loss and average across channels, yielding a remarkably low MSE of 4.93×10^{-5}. Second, we recompose the reconstructed channels into a 6-channel image and apply the same RGB conversion used in the baseline, resulting in an MSE of 4.06×10^{-4}. As shown in Table 6, both variants outperform the RGB baseline, demonstrating that organelle-aware processing enables better use of the pretrained VAE and leads to consistently improved fidelity—even under the same evaluation transformation.

Virtual Instrument. We repeat the model training and morphological analysis, leaving out perturbation 1137 on HUVEC from the training set. We selected this particular combination as this is the most frequent cell type in the dataset and, of the four most frequent perturbations, the one with the largest CATE. In other words, MorphGen has seen many images of this cell type (albeit without this perturbation), and this perturbation, which has a strong effect, was applied to many other samples from other cell types. Perhaps unsurprisingly, we observe no significant performance drop on the held-out group, with FID (38.14 vs. 38.07) and ΔCATE (0.46 vs. 0.43) remaining nearly unchanged. As a result, compositional generalization emerges naturally from the diversity of the training set.

D. Organelle Specific Full Results

Table 7 reports the same metrics as Table 2, with the addition of 95% confidence intervals.

Table 7: FID and KID scores (mean $\pm$ 95% CI) for 50 random perturbations across all cell types. Our method supports generation for all four cell types (HEPG2, HUVEC, RPE, U2OS) and provides channel-wise control.

Channel	FID $\downarrow$	KID $\downarrow$
RGB	50.2 ± 1.3	0.0082 ± 0.0003
Nucleus	27.6 ± 0.6	0.0101 ± 0.0008
ER	48.1 ± 0.4	0.0116 ± 0.0008
Actin	57.6 ± 1.2	0.0155 ± 0.0003
Cyto	49.6 ± 0.4	0.0132 ± 0.0005
Nucleolus	43.6 ± 0.4	0.0123 ± 0.0009
Mito	59.0 ± 2.3	0.0121 ± 0.0018

E. Cell-Type-Specific CATE and Visualizations

Table 8: Conditional Average Treatment Effect (CATE) between control (1138) and perturbed samples, reported per cell type.

Cell Type	p1138 vs p1137			p1138 vs p1108			p1138 vs p1124		
	$\text{CATE}_{\text{real}}$	CATE_{gen}	ΔCATE	$\text{CATE}_{\text{real}}$	CATE_{gen}	ΔCATE	$\text{CATE}_{\text{real}}$	CATE_{gen}	ΔCATE
HEPG2	1.07	1.06	0.01	1.19	0.48	0.71	1.27	0.98	0.29
HUVEC	7.85	7.41	0.44	0.44	0.38	0.06	2.13	2.31	0.18
RPE	1.28	1.09	0.19	1.00	0.65	0.35	1.04	0.83	0.21
U2OS	3.53	2.77	0.76	0.38	0.34	0.04	3.46	2.42	1.04

Table 8 shows the Conditional Average Treatment Effect (CATE) between control (p1138) and perturbed samples, computed using real and generated images across different cell types. Results indicate that images generated by MorphGen preserve treatment-specific cellular features with high fidelity, closely mirroring those from real images. Unsurprisingly, the consistency is strongest for HUVEC cells, likely due to their higher representation in the dataset. The highest treatment effect (i.e., deviation from p1138) in real samples is observed for HUVEC under treatment p1137, with a CATE of 7.85. MorphGen-generated images closely match this effect with a CATE of 7.41, demonstrating consistency even in cases of strong perturbation response. Overall, the closeness of real and generated CATE values suggests that MorphGen-generated images can support accurate downstream analysis.

Real vs Generated

Figure 4 presents PCA visualizations of OpenPhenom (Kraus et al., 2024) features for four representative perturbations – 1108, 1124, 1137 and the control 1138 – arranged in rows. The left column compares real and generated samples by coloring the points by image type. The strong overlap between real and generated distributions across all perturbations indicates that the generated images faithfully reproduce the morphological feature space of the real data. Whereas the right column colors the same embeddings by cell type, revealing clear separability across cell types. This suggest that OpenPhenom features, even when derived from generated images, retain meaningful cell-type structure and can support downstream analyses such as classification. Together, these results demonstrate that MorphGen produces high-quality, morphologically accurate samples that preserve both perturbation effects and intrinsic cell-type differences.

Figures 5, 6, 7, and 8 show PCA visualizations at the cell-type level (HEPG2, HUVEC, RPE, and U2OS, respectively), allowing a qualitative assessment of color separation across different perturbations. Overall, the results demonstrate that features extracted from MorphGen-generated images exhibit: (i) strong alignment with real features, making them visually indistinguishable, and (ii) clear separability with respect to both cell type and perturbation—the two generative factors we explicitly control.

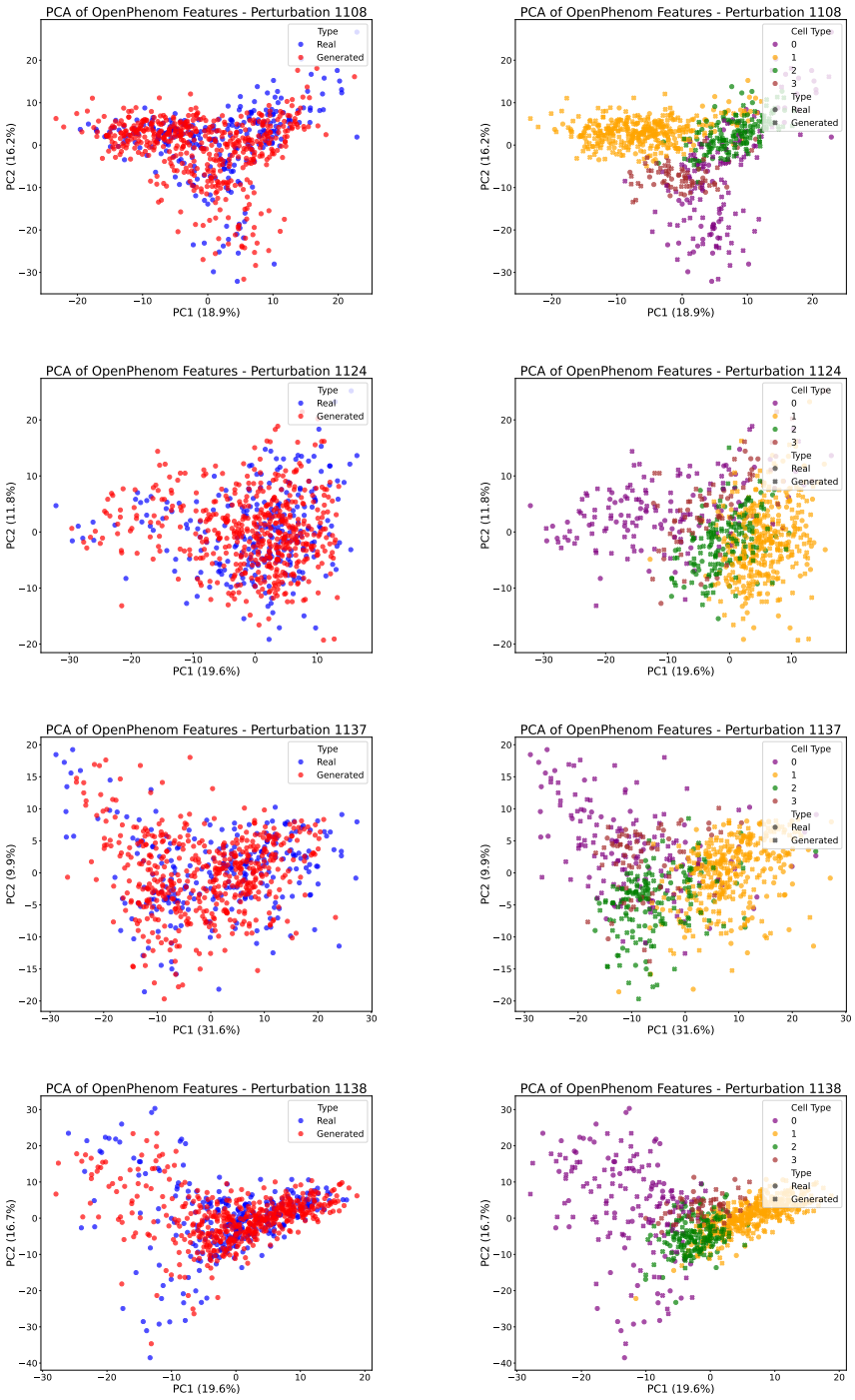
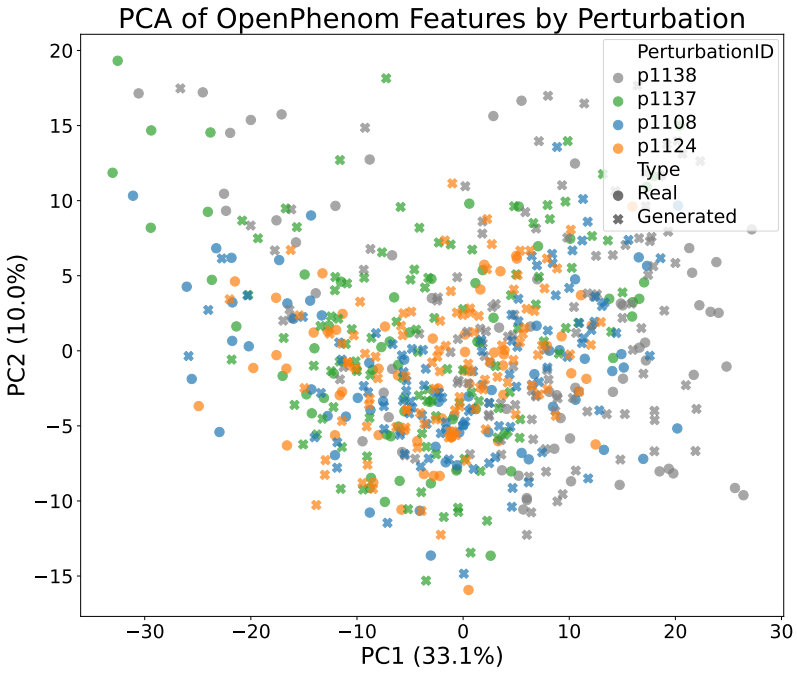
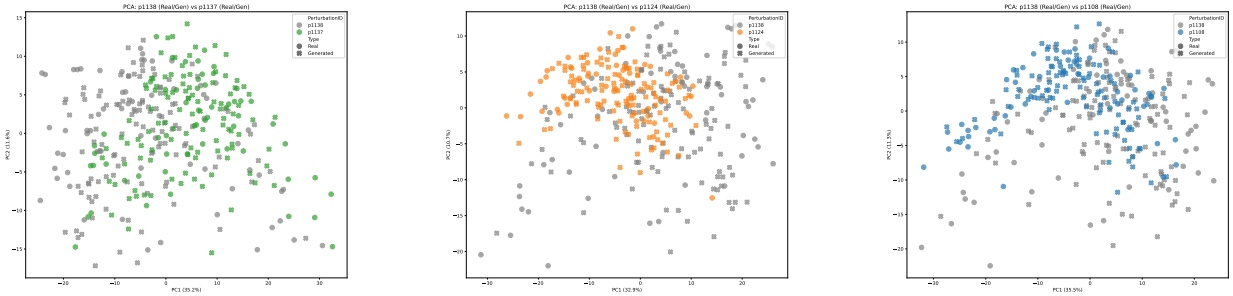


Figure 4: PCA visualizations of OpenPhenom features for four perturbations (rows: 1108, 1124, 1137, and control 1138). Each row compares real and generated embeddings for a single perturbation. *Left column:* points are colored by image type (real vs. generated), revealing strong overlap—indicating that generated images closely match the distribution of real samples. *Right column:* the same embeddings are now colored by cell type, showing that cell-type-specific structure is preserved in both real and generated data. Together, these plots demonstrate that our model produces high-quality, morphologically faithful samples that capture both perturbation effects and intrinsic cell type differences.

Perturbation Effects Visualizations HEPG2



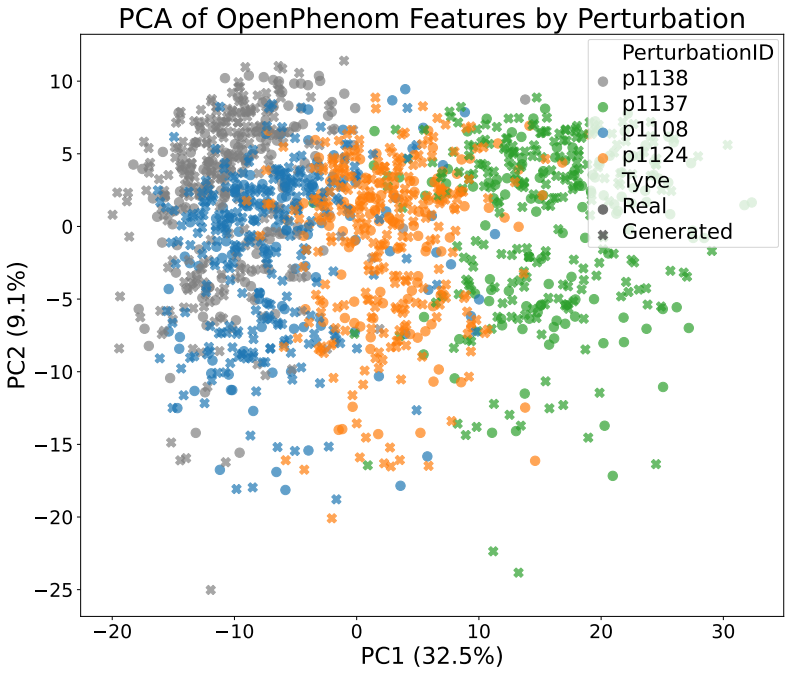
(a) PCA visualization across all four perturbations, including the control (1138).



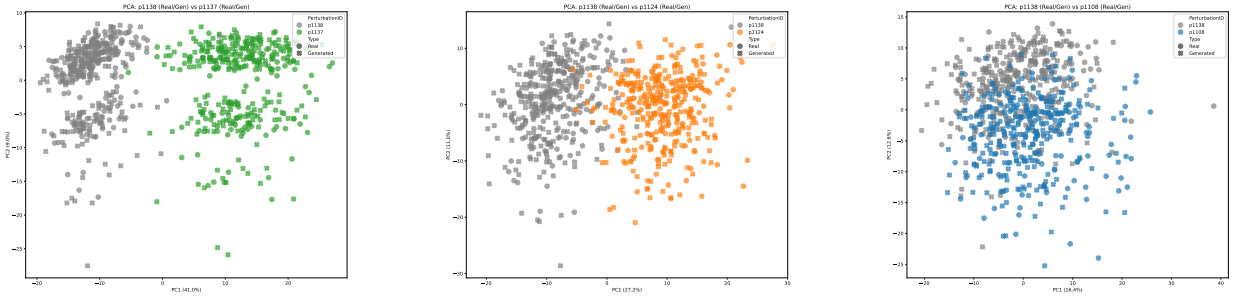
1138 vs 1137 | 1138 vs 1124 | 1138 vs 1108

Figure 5: PCA projections of phenotypic embeddings of HEPG2 cells. The top panel shows global variation across all perturbations. The bottom panels show focused pairwise comparisons between the control (1138) and specific perturbations.

Perturbation Effects Visualizatons HUVEC



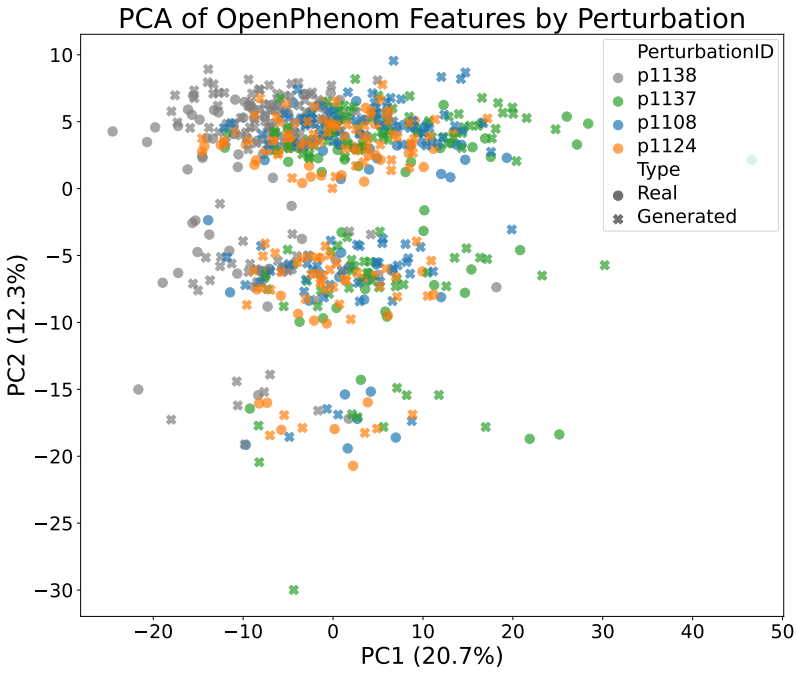
(a) PCA visualization across all four perturbations, including the control (1138).



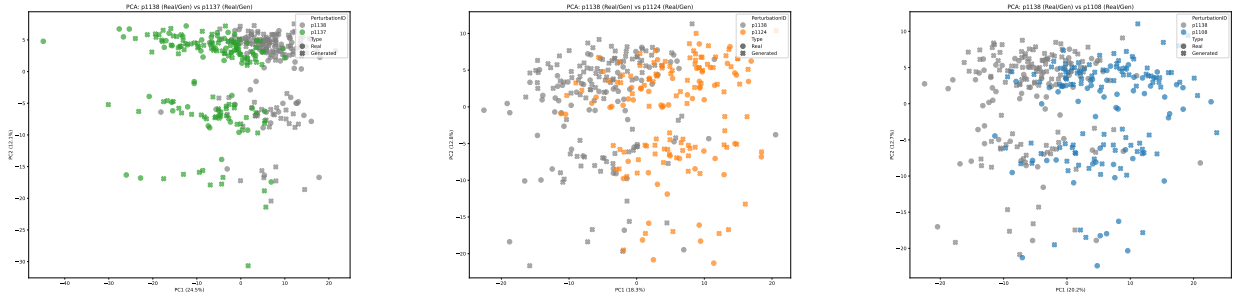
1138 vs 1137 | 1138 vs 1124 | 1138 vs 1108

Figure 6: PCA projections of phenotypic embeddings of HUVEC cells. The top panel shows global variation across all perturbations. The bottom panels show focused pairwise comparisons between the control (1138) and specific perturbations.

Perturbation Effects Visualizations RPE



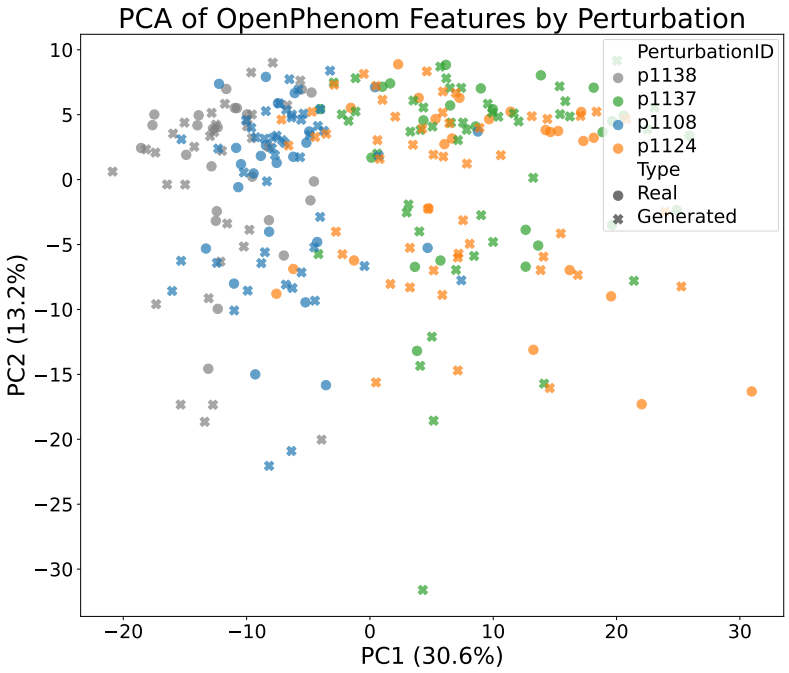
(a) PCA visualization across all four perturbations, including the control (1138).



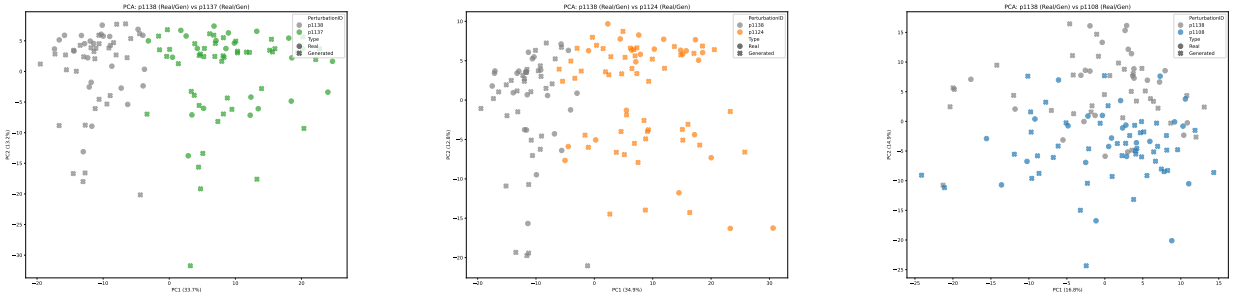
1138 vs 1137 | 1138 vs 1124 | 1138 vs 1108

Figure 7: PCA projections of phenotypic embeddings of RPE cells. The top panel shows global variation across all perturbations. The bottom panels show focused pairwise comparisons between the control (1138) and specific perturbations.

Perturbation Effects Visualizations U2OS



(a) PCA visualization across all four perturbations, including the control (1138).



1138 vs 1137 | 1138 vs 1124 | 1138 vs 1108

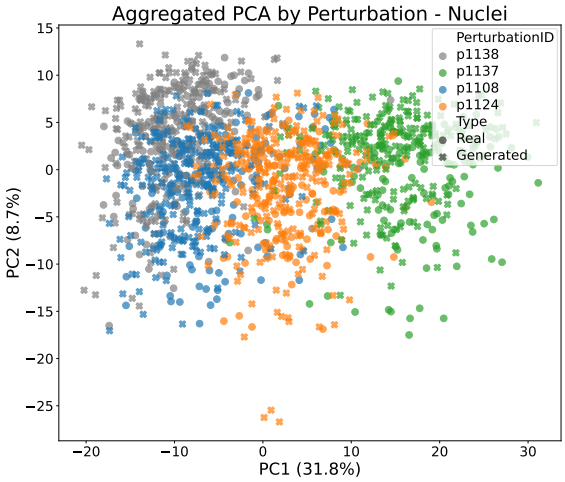
Figure 8: PCA projections of phenotypic embeddings of U2OS cells. The top panel shows global variation across all perturbations. The bottom panels show focused pairwise comparisons between the control (1138) and specific perturbations.

F. Organelle-Specific CATE and Visualizations

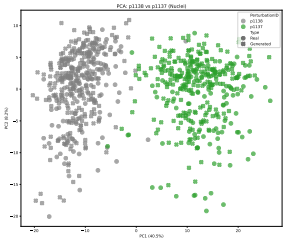
Table 9 reports organelle-specific CATEs, measuring the deviation of perturbations p1108, p1124, and p1137 from the control p1138. Notably, our estimates closely match the real values even at the organelle level, suggesting that MorphGen can accurately capture organelle-specific response patterns. Figure 9 further illustrates this by showing PCA visualizations of Nuclei responses across different perturbations.

Table 9: Conditional Average Treatment Effect (CATE) between control (1138) and perturbed HUVEC cells, reported per organelle.

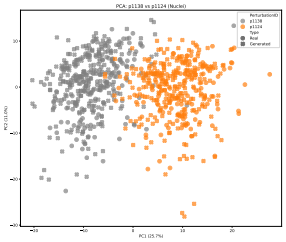
Organelle	**p1138 vs p1137**			**p1138 vs p1108**			**p1138 vs p1124**		
	$\text{CATE}_{\text{real}}$	CATE_{gen}	ΔCATE	$\text{CATE}_{\text{real}}$	CATE_{gen}	ΔCATE	$\text{CATE}_{\text{real}}$	CATE_{gen}	ΔCATE
Nuclei	9.47	9.50	0.03	0.69	0.59	0.10	2.86	2.97	0.11
ER	9.46	9.54	0.08	0.69	0.60	0.09	2.85	2.97	0.12
Actin	9.49	9.50	0.01	0.69	0.59	0.10	2.86	2.97	0.11
Nucleoli	9.49	9.51	0.02	0.69	0.59	0.10	2.85	2.96	0.11
Mitochandria	9.54	9.51	0.03	0.70	0.59	0.11	2.86	2.97	0.11
Golgi	9.49	9.52	0.03	0.69	0.60	0.09	2.86	2.97	0.11



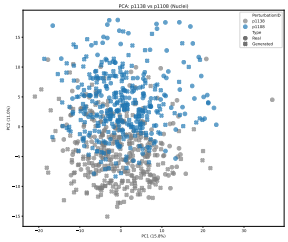
(a) PCA visualization across all four perturbations, including the control (1138).



1138 vs 1137



1138 vs 1124



1138 vs 1108

Figure 9: PCA projections of Nuclei embeddings of HUVEC cells. The top panel shows global variation across all perturbations. The bottom panels show focused pairwise comparisons between the control (1138) and specific perturbations.