Compressing Biology: Evaluating the Stable Diffusion VAE for Phenotypic Drug Discovery

Télio Cropsal
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
Gothenburg, SE
telio@chalmers.se

Rocío Mercado
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
Gothenburg, SE
rocio.mercado@chalmers.se

Abstract

High-throughput phenotypic screens generate vast microscopy image datasets that push the limits of generative models due to their large dimensionality. Despite the growing popularity of general-purpose models trained on natural images for microscopy data analysis, their suitability in this domain has not been quantitatively demonstrated. We present the first systematic evaluation of Stable Diffusion's variational autoencoder (SD-VAE) for reconstructing Cell Painting images, assessing performance across a large dataset with diverse molecular perturbations and cell types. We find that SD-VAE reconstructions preserve phenotypic signals with minimal loss, supporting its use in microscopy workflows. To benchmark reconstruction quality, we compare pixel-level, embedding-based, latent-space, and retrieval-based metrics for a biologically informed evaluation. We show that general-purpose feature extractors like InceptionV3 match or surpass publicly available bespoke models in retrieval tasks, simplifying future pipelines. Our findings offer practical guidelines for evaluating generative models on microscopy data and support the use of off-the-shelf models in phenotypic drug discovery.

1 Introduction

Phenotypic drug discovery is a strategy in drug development that identifies drug candidates by directly observing their effects in biological systems without requiring prior knowledge of molecular targets [30]. By focusing on measurable phenotypic changes induced by molecular perturbations, this approach has historically led to the discovery of several clinically relevant drugs, such as the anti-malarial artemisinin [18]. Recent advances in high-throughput microscopy, e.g., Cell Painting [3], have accelerated phenotypic screening pipelines [24]. In Cell Painting, cells are stained with multiple fluorescent dyes that mark distinct subcellular components, enabling the capture of rich morphological profiles via automated, low-cost fluorescence microscopy. A single lab can thus yield millions of high-res, multi-channel images under diverse conditions. These images can be processed with tools such as CellProfiler [28] to extract thousands of morphological features per condition. However, the sheer dimensionality of the resulting profiles pose significant analytical challenges, motivating the development of methods that can uncover subtle phenotypic patterns at scale.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

Deep learning (DL) methods have become popular tools for addressing these challenges, initially via representation learning methods to extract meaningful features from raw images [11, 20, 27]. More recently, the success of generative models has inspired efforts to simulate Cell Painting images, with the aim of reducing experimental cost and enabling virtual screening [7]. Several studies have shown promising results in generating realistic microscopy images conditioned on molecular [31] or genetic [19] perturbations. Despite this progress, generative modeling of microscopy data remains computationally challenging due to its high dimensionality. Direct pixel-space generation is costly, leading many to restrict image generation tasks to small crops centered around individual nuclei [2, 21, 33] or to adopt latent diffusion approaches [19, 22]. Latent diffusion models, exemplified by Stable Diffusion (SD) [26], tackle complexity by first mapping images to a compressed latent representation using a variational autoencoder (VAE) [13], then performing diffusion-based generative modeling within this lower-dimensional space. At inference, images are generated in latent space and decoded back to image space, with the VAE acting as a bottleneck for the final reconstruction quality.

Reliable generation of microscopy images is critical for downstream biological interpretation, and the SD-VAE is increasingly used for this purpose [19, 22]. However, the reconstruction quality of SD-VAE-generated images has not been systematically evaluated. This raises concerns about whether meaningful biological information may be lost during the encoding-decoding process, especially when applied to out-of-distribution microscopy data such as Cell Painting images. To address this gap, this work evaluates the reconstruction fidelity of SD-VAE on Cell Painting images using a recently established benchmark.

Our main contributions address this gap and are as follows:

- **First systematic evaluation of SD-VAE on microscopy data:** We evaluated SD-VAE, trained primarily on natural images, on >1M Cell Painting crops spanning two cell types (A549 and U2OS) and diverse perturbations. We demonstrate that SD-VAE reconstructions retain phenotypic signals with minimal degradation, validating its use in microscopy image generation.
- **A general evaluation framework for generative models of microscopy images:** We systematically compare pixel-level metrics, e.g., mean absolute error (MAE) and earth mover's distance (EMD), with feature-space and latent-space metrics, e.g., Kullback-Leibler divergence (KLD) and Fréchet Inception distance (FID), as well as retrieval-based evaluations, presenting a robust framework that researchers can model future validation studies on.
- **General-purpose feature extractors rival domain-specific models:** We find that InceptionV3 embeddings match or exceed those from the publicly-available microscopy-specialized OpenPhenom model on biologically relevant retrieval tasks, suggesting that general-purpose feature extractors may be sufficient to capture subtle phenotypic variations.

2 Method

Our benchmarking pipeline uses multiple quantitative metrics to evaluate the effectiveness of SD-VAE in compressing and reconstructing microscopy data. We use two distinct pre-trained models for image featurization: InceptionV3 [29], trained on natural images, and OpenPhenom [15, 23], trained specifically on Cell Painting images. We do not compare the feature extractors with CellProfiler [28] due to its high computational cost and the complexity of integrating it into scalable, automated GPU-based workflows. Instead, we rely on deep learning-based models, which are better suited to our infrastructure and more representative of how generative models are typically deployed in production settings. We assess reconstruction quality as described in Section 2.3.

2.1 Pre-trained models

Stable Diffusion VAE We evaluate the VAE associated with the model *stable-diffusion-v1-4* [6, 25], which was trained on natural images of varying dimensions (256x256 and 512x512) from the *laion2B-en* dataset. This autoencoder employs a relative downsampling factor of 8. For example, an RGB image with a resolution of (3x256x256) would be mapped to a latent tensor of shape (4x32x32). Throughout all experiments the VAE weights are kept frozen and used only for encoding and decoding.

InceptionV3 As a baseline feature extraction method we use the *torchvision* implementation of InceptionV3 [29], which has been pretrained on the ImageNet dataset [8]. To evaluate the capabilities of models for image generation, the Fréchet distance is typically applied to features extracted from

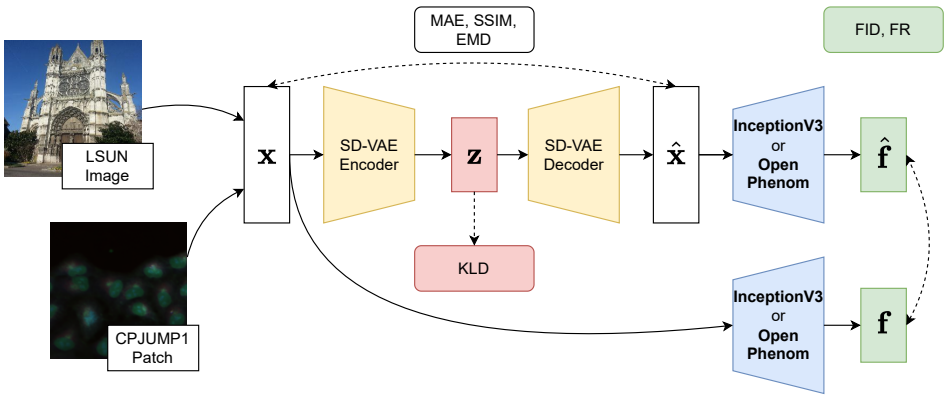

Figure 1: Overview of our evaluation pipeline. Input images, including Cell Painting and natural images, are encoded into the latent space using the Stable Diffusion VAE (SD-VAE). These latent representations are decoded to reconstruct the images. Original and reconstructed images are compared using channel-wise metrics: mean absolute error (MAE), structural similarity index measure (SSIM), and Earth mover's distance (EMD). Both sets of real and reconstructed images are passed through InceptionV3 and OpenPhenom to extract feature embeddings, which are used to compute the Fréchet Inception distance (FID) and fraction retrieved (FR) of perturbations against negative controls. Latent vectors are further used to compute the Kullback-Leibler divergence (KLD).

the layer just before the final classification layer of an InceptionV3 network. The empirical Fréchet distance computed between the means and covariances of these features following real and generated distributions is known as the Fréchet Inception distance (FID) [10] (see Appendix A.2).

OpenPhenom As an alternative to CellProfiler, we utilize a pre-trained channel-agnostic masked autoencoder (CA-MAE) called OpenPhenom, which employs a ViT-S/16 encoder backbone to generate embeddings for Cell Painting images [15]. OpenPhenom is a fully open-access model available on HuggingFace [23], trained on >3M 256×256 image crops from publicly available Cell Painting datasets with genetic perturbations, including RxRx3 (HUVEC cell line) [9] and the JUMP-CP CRISPR and ORF subsets (U2OS cell line) [5]. The model was trained for 100 epochs using a CA-MAE architecture with a 25M parameter ViT-S/16 encoder and six dedicated decoders (one per channel). The training channels include *mitochondria*; *DNA* (nucleus); *RNA*; *endoplasmic reticulum (ER)*; *actin, Golgi and plasma membrane (AGP)*; and a non-fluorescent Brightfield channel. During training, each 16×16 patch from each channel is tokenized independently, resulting in 1,536 tokens for a 6-channel image, with 384 unmasked tokens visible under a 75% masking ratio. OpenPhenom can generate embeddings either for the entire image or per individual channel.

2.2 Data

We evaluate the performance of the SD-VAE on one out-of-distribution Cell Painting dataset and, as a control, one in-distribution natural image dataset. For data preprocessing details, see Appendix A.3.

CPJUMP1 This dataset [4] includes replicated plates containing both chemical and genetic perturbations with known mechanisms of action or molecular targets. These replicates are tested under different experimental conditions, including two cell lines (A549 and U2OS) and two exposure durations, specifically 24 hours and 48 hours for compound treatments, representing short and long time points. For this study, we used a subset of the dataset that includes only chemical perturbations with 100% cell seeding and parental cell lines (66,048 center-cropped 1024x1024 images among 16 plates exposed to 307 unique perturbations, including DMSO controls). Each image contains 5 channels, corresponding in this dataset to fluorescent stains of the mitochondria (*Mito*), actin, Golgi, and plasma membrane (*AGP*), nucleoli and cytoplasmic RNA (*RNA*), endoplasmic reticulum (*ER*), and nuclear DNA (*DNA*).

LSUN As a control, we also assess the performance of SD-VAE on high-res natural images which closely resemble the VAE's training data. Specifically, we utilize two subsets from the LSUN [32]

dataset: the classroom subset, comprising 166,419 images, and the outdoor church subset, containing 126,200 images. All images are already resized so that the smaller dimension is 256 pixels. Images consist of the three standard RGB channels. Our evaluation on this dataset uses the same set of metrics for consistency.

2.3 Evaluation

Images or patches with a 256×256 pixel resolution are first passed through the encoder and decoder of the SD-VAE. The resulting reconstructions are then compared to the original images to assess reconstruction quality. Moreover, both the original and reconstructed images are processed through two pre-trained networks, InceptionV3 and OpenPhenom, for independent feature extraction to assess the reconstruction of morphological profiles via either scheme. Details and equations for all metrics are provided in Appendix A.2.

Channel- and distribution-wise We compute the mean absolute error (MAE), structural similarity index measure (SSIM), and Earth mover's distance (EMD) between each channel of the real and reconstructed images. While MAE provides a straightforward pixel-wise error, SSIM captures perceptual differences by considering structural information, and EMD evaluates how closely the distributions of pixel intensities match between the original and reconstructed images. Conversely, the Fréchet Inception distance [10] is a widely used distribution-based metric for evaluating image generation models. It measures the difference between the distributions of real and generated images by comparing the means and covariances of features extracted from the Inception network.

Regularized latent space To complement the aforementioned metrics and evaluate the quality of the learned latent space by the SD-VAE, we evaluate the Kullback-Leibler divergence (KLD) between the samples in the latent space and a standard multivariate Gaussian distribution to assess how well the latent space is regularized, indicating how easily it can be learned, e.g., by a diffusion model.

Information retrieval To better reflect the practical application of generative models in phenotypic drug discovery, we follow the evaluation procedure described by Chandrasekaran et al. [4] and Kalinin et al. [12]. This approach evaluates the quality of learned embeddings through a retrieval task that identifies replicates of the same perturbation, target, or mechanism of action (MoA). Here we compare the InceptionV3 network and the pre-trained masked autoencoder OpenPhenom. The retrieval task is performed separately on embeddings obtained from real and reconstructed images. Plate-specific batch effects are accounted for and mitigated through post-processing of the features extracted from InceptionV3 or OpenPhenom using negative control wells, as detailed in Appendix A.4. Our evaluation focuses solely on phenotypic activity, as it serves as a challenging sanity check for downstream applications. Specifically, we assess whether replicate profiles for a given perturbation can be distinguished from replicate profiles under control conditions, using negative controls as the reference. Hence, the fraction retrieved (FR) metric quantifies the proportion of perturbations that can be reliably distinguished from negative controls. For this purpose, we employ the *copairs* library [12], specifically designed for retrieval-based analysis. Features extracted from each pre-trained network are aggregated across imaging sites to produce a single feature vector per well. Features extracted from InceptionV3 are concatenated across the two distinct 3-channel input images, resulting in a per-sample dimensionality of $2 \times 2048 = 4096$. Similarly, OpenPhenom features are concatenated along the channel axis, yielding a per-sample dimensionality of $5 \times 384 = 1920$. See Appendix A.4 for key post-processing details.

3 Results

3.1 SD-VAE reconstructs images well in terms of standard reconstruction metrics

Cell Painting images show a lower MAE between original and reconstructed samples (Figure 2), indicating that SD-VAE effectively reconstructs them. This aligns with expectations, as Cell Painting images typically contain simpler and more structured visual patterns than natural images, with certain channels often displaying a relatively consistent background. Supporting metrics such as EMD, FID, and SSIM also show similar value ranges across both image types (Figures 2–4), reinforcing that reconstructed Cell Painting images are visually faithful to the originals.

3.2 Biological signal preserved after SD-VAE reconstruction

While metrics like MAE confirm better pixel-level reconstruction of cell images than natural images, they do not capture the reconstruction of relevant biological features. Instead, metrics like the fraction retrieved (FR) can help us infer if biological signal is preserved following application of the SD-VAE. Notably, when evaluating the FR across different cell lines and time points (Table 1), we found that the reconstructed images from SD-VAE did not lead to a significant drop in FR, even demonstrating a slight increase in many cases. This suggests that, despite the reconstruction process, the biological signal remains sufficiently intact to distinguish between negative controls and perturbations.

Table 1: Fraction retrieved (FR) across cell types and time points, with per-well median aggregation and plate-level mean scaling of DL features. The reported values are averaged over three independent runs of the complete pipeline (SD-VAE and information retrieval). The largest std. dev. is $\sigma = 0.008$. Experiments with original images by design show no variation ($\sigma = 0$). The best value for each cell line and time point is shown in bold. The bottom row compares performance on the same task for a traditional feature extraction method.

		A549		U2OS	
Features	**Data**	**24h**	**48h**	**24h**	**48h**
OpenPhenom	Original	0.722	0.882	0.817	0.660
OpenPhenom	SD-VAE	0.729	0.879	0.836	0.697
InceptionV3	Original	0.873	**0.961**	0.837	**0.837**
InceptionV3	SD-VAE	**0.906**	0.951	**0.847**	**0.837**
CellProfiler [4]	Original	0.761	0.954	0.775	0.663

3.3 KLD suggests microscopy latents are less regularized than those of natural images

We observe that Cell Painting images exhibit a higher Kullback–Leibler divergence (KLD) between their latent representations and an isotropic Gaussian prior, compared to natural images (Figure 3). This indicates that the latent space for microscopy data is less regularized and the encoded representations deviate more from the prior distribution. In contrast, natural images produce latent vectors that are closer to the prior. This difference suggests that it is more difficult for the model to compress the complex, biologically rich content of Cell Painting images into a smooth, well-structured latent space. As a result, tasks that rely on latent representations may be affected by this reduced regularization, e.g., this may complicate the training of downstream latent diffusion models.

4 Discussion

While metrics like MAE, SSIM, EMD, and FID are useful for assessing low-level similarity, they provide limited insight into the preservation of biological signal. This motivates the use of more biologically grounded and interpretable evaluation strategies, such as measuring the FR of perturbations against negative controls, to assess whether reconstructions retain relevant phenotypic information.

Our experiments show that general-purpose feature extractors, such as InceptionV3, can perform on par with, and in some cases better than, domain-specific models like OpenPhenom in tailored retrieval tasks. This suggests that models pre-trained on natural images may be sufficiently effective for evaluating generative models in the context of phenotypic drug discovery, reducing the need for specialized feature extractors. The relatively weaker performance of OpenPhenom relative to InceptionV3 can be attributed to several factors. First, there is a discrepancy between the training and inference conditions: during inference, only five channels are provided while the Brightfield channel (used during OpenPhenom training) is omitted. Second, the OpenPhenom model used here is a comparatively small model trained on a limited dataset, unlike the other larger CA-MAE variants which are unfortunately not publicly available. Overall, our results support the use of FID as a practical and reliable metric during model development and evaluation, as evidenced by its alignment with the FR and the demonstrated effectiveness of Inception features in the retrieval task.

Note that in this study, we deliberately avoid evaluating the SD-VAE latent space; this is because evaluating the latent space in greater depth beyond KLD would require specialized methods. In this work, we instead rely on existing models (InceptionV3 and OpenPhenom) to evaluate the reconstructed images rather than the latent space. This setup better reflects future use cases of generative models, where new samples are generated and assessed by surrogate models in an automated fashion. Additionally, it has been shown that the latent space of SD-VAE is not strongly semantically regularized; rather, it serves as a compressed representation of the original image, removing redundant information while preserving spatial structure [14]. We leave for future work a comparison between SD-VAE and other dimensionality reduction techniques, since it would also be necessary to demonstrate that the resulting latent space is suitable for generating Cell Painting images. This has already been shown with MorphoDiff [19], although their evaluation did not isolate the performance of SD-VAE. For similar reasons, we have not fine-tuned SD-VAE, as the straightforward approach is known to be ineffective [17].

There are a few additional limitations to our approach. First, our dataset contains five distinct channels, but both the VAE and the InceptionV3 model require 3-channel inputs. To address this, we duplicated one of the channels to create two separate 3-channel combinations, allowing us to use all five channels while maintaining compatibility with the models. Although effective, this approach is somewhat arbitrary and may not be optimal. Future work could investigate more systematic strategies for channel grouping or selection. Second, there is the risk of data leakage. Using negative controls in the post-processing pipeline is standard practice for limiting batch effects. Since FR involves distinguishing perturbations from negative controls, this may be a form of data leakage. Moreover, we observed an increase in the FR metric across most experiments involving reconstructed images. This trend may suggest a denoising effect, as the reconstruction process could be removing noise or artifacts in ways that improve FR performance. While we acknowledge this potential limitation, addressing it in depth is also left for future work.

5 Conclusion

When working with high-dimensional Cell Painting images, SD-VAE appears to be a promising approach for image generation that mostly preserves biological signal while reducing the dimensionality of the data. This is critical because the generated images or even latents are typically be used as input for downstream models that aim to identify meaningful patterns in this high-dimensional data. With this work, we provide a framework to ensure that the VAE, or any other generative model, can be adequately integrated into the overall workflow without significant degradation of biological signal. Our work further supports the use of SD-VAE and general metrics like FID in Cell Painting analysis workflows without the need for specialized training of bespoke models.

Acknowledgments

TC and RM acknowledge the funding provided by the Wallenberg AI, Autonomous Systems, and Software Program (WASP), supported by the Knut and Alice Wallenberg Foundation. The computations and data storage were enabled by resources provided by Chalmers e-Commons at Chalmers. The computations and data storage were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Code and Data Availability

For implementation details and source code, please refer to our anonymized GitHub repository: `compressing-biology`. Pre-trained models and datasets used can be downloaded from:

- SD-VAE: huggingface.co/CompVis/stable-diffusion-v1-4
- InceptionV3: docs.pytorch.org/vision/main/models/inception.html
- OpenPhenom: huggingface.co/recursionpharma/OpenPhenom
- CPJUMP1: `cellpainting-gallery.s3.amazonaws.com`
- LSUN: docs.pytorch.org/vision/main/generated/torchvision.datasets.LSUN.html

References

[1] D. M. J. Ando, Cory Y. McLean, and Marc Berndl. Improving phenotypic measurements in high-content imaging screens. *bioRxiv*, 2017. URL https://api.semanticscholar.org/CorpusID:26552204.

[2] Anis Bourou, Thomas Boyer, Marzieh Gheisari, Kévin Daupin, Véronique Dubreuil, Aurélie De Thonel, Valérie Mezger, and Auguste Genovesio. PhenDiff: Revealing subtle phenotypes with diffusion models in real images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 358–367. Springer, 2024.

[3] 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. *Nature Protocols*, 11(9):1757–1774, 2016.

[4] Srinivas Niranj Chandrasekaran, Beth A. Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, John Arevalo, Juan C. Caicedo, Daniel Kuhn, Desiree Hernandez, Jim Berstler, Hamdah Shafqat-Abbasi, David E. Root, Sussane Swalley, Shantanu Singh, and Anne E Carpenter. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21:1114 – 1121, 2022. URL https://api.semanticscholar.org/CorpusID:237250709.

[5] Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, pages 2023–03, 2023.

[6] CompVis. Stable Diffusion v1-4. https://huggingface.co/CompVis/stable-diffusion-v1-4, 2022. Accessed: 2025-08-06.

[7] Jan Oscar Cross-Zamirski, Praveen Anand, Guy Williams, Elizabeth Mouchet, Yinhai Wang, and Carola-Bibiane Schönlieb. Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3800–3809, 2023.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[9] Marta M. Fay, Oren Z. Kraus, Mason L. Victors, Lakshmanan Arumugam, Kamal Vuggumudi, John Urbanik, Kyle Hansen, Safiye Celik, Nico Cernek, Ganesh Jagannathan, Jordan Christensen, Berton A. Earnshaw, Imran S. Haque, and Ben Mabey. RxRx3: Phenomics map of biology. *bioRxiv*, 2023. URL https://api.semanticscholar.org/CorpusID:256699099.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv*, 2018. URL https://arxiv.org/abs/1706.08500.

[11] Maria Hofmarcher, Eva Rumetshofer, Djork-Arné Clevert, Sepp Hochreiter, and Günter Klambauer. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *Journal of Chemical Information and Modeling*, 59(3):1163–1171, 2019. doi: 10.1021/acs.jcim.8b00670. URL https://pubs.acs.org/doi/10.1021/acs.jcim.8b00670.

[12] Alexandr A. Kalinin, John Arevalo, Loan Vulliard, Erik Serrano, Hillary Tsang, Michael Bornholdt, Bartek Rajwa, A.E. Carpenter, Gregory P. Way, and Shantanu Singh. A versatile information retrieval framework for evaluating profile strength and similarity. *Nature Communications*, 16, 2024. URL https://api.semanticscholar.org/CorpusID:268930261.

[13] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019. doi: 10.1561/2200000056. URL https://www.nowpublishers.com/article/Details/MAL-056.

[14] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling, 2025. URL https://arxiv.org/abs/2502.09509.

[15] 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. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11757–11768, 2024. URL https://api.semanticscholar.org/CorpusID:269157464.

[16] Oren Kraus, Federico Comitani, John Urbanik, Kian Kenyon-Dean, Lakshmanan Arumugam, Saber Saberian, Cas Wognum, Safiye Celik, and Imran S. Haque. RxRx3-core: Benchmarking drug-target interactions in high-content microscopy. *arXiv*, 2025. URL https://arxiv.org/abs/2503.20158.

[17] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers, 2025. URL https://arxiv.org/abs/2504.10483.

[18] Louis H Miller and Xinzhuan Su. Artemisinin: discovery from the Chinese herbal garden. *Cell*, 146(6): 855–858, 2011.

[19] Zeinab Navidi, Jun Ma, Esteban A Miglietta, Le Liu, Anne E Carpenter, Beth A Cimini, Benjamin Haibe-Kains, and Bo Wang. MorphoDiff: Cellular morphology painting with diffusion models. *bioRxiv*, pages 2024–12, 2024.

[20] Hieu Nguyen, Hugo Sanchez, and Anne E. Carpenter. Molecule-morphology contrastive pretraining for enhanced phenotypic predictions. *iScience*, 31(12):106892, 2024. doi: 10.1016/j.isci.2024.106892. URL https://www.cell.com/iscience/fulltext/S2589-0042(24)02659-2.

[21] Alessandro Palma, Fabian J Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16(1):505, 2025.

[22] Giorgos Papanastasiou, Pedro P Sanchez, Argyrios Christodoulidis, Guang Yang, and Walter Hugo Lopez Pinaya. Confounder-aware foundation modeling for accurate phenotype profiling in cell imaging. *bioRxiv*, pages 2024–12, 2024.

[23] Recursion Pharmaceuticals. OpenPhenom: Channel-agnostic masked autoencoder for Cell Painting images. https://huggingface.co/recursionpharma/OpenPhenom, 2024. Accessed: 2025-08-06.

[24] Jonne Rietdijk, Marianna Tampere, Aleksandra Pettke, Polina Georgiev, Maris Lapins, Ulrika Warpman-Berglund, Ola Spjuth, Marjo-Riitta Puumalainen, and Jordi Carreras-Puigvert. A phenomics approach for antiviral drug discovery. *BMC Biology*, 19(1):156, 2021.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv*, 2022. URL https://arxiv.org/abs/2112.10752.

[27] Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023.

[28] David R Stirling, Madison J Swain-Bowden, Alice M Lucas, Anne E Carpenter, Beth A Cimini, and Allen Goodman. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, 22:1–11, 2021.

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *arXiv*, 2015. URL https://arxiv.org/abs/1512.00567.

[30] Fabien Vincent, Arsenio Nueda, Jonathan Lee, Monica Schenone, Marco Prunotto, and Mark Mercola. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, 21(12):899–914, 2022.

[31] Karren Yang, Samuel Goldman, Wengong Jin, Alex X. Lu, Regina Barzilay, Tommi Jaakkola, and Caroline Uhler. Mol2Image: Improved conditional flow models for molecule to image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6687–6696, 2021. doi: 10.1109/CVPR46437.2021.00662. URL https://openaccess.thecvf.com/content/CVPR2021/papers/Yang_Mol2Image_Improved_Conditional_Flow_Models_for_Molecule_to_Image_Synthesis_CVPR_2021_paper.pdf.

[32] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv*, 2016. URL https://arxiv.org/abs/1506.03365.

[33] Yuhui Zhang, Yuchang Su, Chenyu Wang, Tianhong Li, Zoe Wefers, Jeffrey Nirschl, James Burgess, Daisy Ding, Alejandro Lozano, Emma Lundberg, and Serena Yeung-Levy. CellFlux: Simulating cellular morphology changes via flow matching. *arXiv*, 2025. URL https://arxiv.org/abs/2502.09775.

A Technical Appendices and Supplementary Material

A.1 Notation

Let $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ denote the ground truth image and $\hat{\mathbf{x}} \in \mathbb{R}^{C \times H \times W}$ the reconstructed image. The following notation is used:

- C: number of channels
- H, W: image height and width
- $N = H \times W$: number of pixels per channel
- $\mu_{\mathbf{x}}, \mu_{\hat{\mathbf{x}}}$: local means
- $\sigma^2_{\mathbf{x}}, \sigma^2_{\hat{\mathbf{x}}}$: local variances
- $\sigma_{\mathbf{x}\hat{\mathbf{x}}}$: local covariance
- C_L: constant to stabilize luminance comparison
- C_C: constant to stabilize contrast and structure comparison
- $\mu_i, \log \sigma_i^2$: mean and log-variance of latent variable z_i
- d: dimensionality of the latent space
- $\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r$: mean and covariance of real image features
- $\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$: mean and covariance of generated image features

A.2 Evaluation Metrics

Below we define the pixel- and distribution-based metrics used in this study:

1. **Mean absolute error (MAE)** The average absolute difference between corresponding pixels:

$$\text{MAE}_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |\mathbf{x}_{c,i,j} - \hat{\mathbf{x}}_{c,i,j}| \tag{1}$$

2. **Structural similarity index (SSIM)** A perceptual similarity measure combining luminance, contrast, and structure:

$$\text{SSIM}_c(\mathbf{x}, \hat{\mathbf{x}}) = \frac{(2\mu_{\mathbf{x}}\mu_{\hat{\mathbf{x}}} + C_L)(2\sigma_{\mathbf{x}\hat{\mathbf{x}}} + C_C)}{(\mu^2_{\mathbf{x}} + \mu^2_{\hat{\mathbf{x}}} + C_L)(\sigma^2_{\mathbf{x}} + \sigma^2_{\hat{\mathbf{x}}} + C_C)} \tag{2}$$

3. **Earth mover's distance (EMD)** The average absolute difference between sorted pixel intensities:

$$\text{EMD}_c = \frac{1}{N} \sum_{k=1}^{N} |\text{sort}(\mathbf{x}_c)_k - \text{sort}(\hat{\mathbf{x}}_c)_k| \tag{3}$$

4. **Kullback–Leibler divergence (KLD)** The divergence between the latent distribution and a standard Gaussian:

$$\text{KL} = \frac{1}{2} \sum_{i=1}^{d} \left(\mu_i^2 + \exp(\log \sigma_i^2) - \log \sigma_i^2 - 1\right) \tag{4}$$

5. **Fréchet inception distance (FID)** The distance between real and generated feature distributions:

$$\text{FID} = \left\|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\right\|^2 + \text{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\left(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g\right)^{1/2}\right) \tag{5}$$

A.3 Data Pre-Processing Details

Images from the CPJUMP1 dataset are processed similar to RxRx3-Core [16]. The provided illumination correction arrays are applied, and images are saved as *uint8* and compressed as PNGs. LSUN images are already optimized for DL pipelines. Images are standardized to 256×256 pixels to ensure consistent resolution: LSUN images are directly resized to the target resolution via interpolation if needed, whereas CPJUMP1 images are divided into 256×256 pixel patches. As SD-VAE and InceptionV3 expect the standard 3-channel (RGB) inputs, the 5-channel Cell Painting images are handled differently: one of the five channels (RNA, selected randomly) is duplicated to create 6 channels. These are then split into two separate 3-channel images following the approach of Papanastasiou et al. [22]. For the OpenPhenom model, images are pre-processed using self-standardization, as recommended by Kraus et al. [15], and kept in the 5-channel format as the model is channel agnostic.

A.4 Data Post-Processing Details

Extracted feature vectors are post-processed using a pipeline inspired by the typical variation normalization (TVN) method [1] recommended by OpenPhenom. Post-processing is done as follows:

- Fit a sequence of preprocessing steps, including scaling, principal component analysis (PCA), and variance thresholding, to all negative control samples, using the highest feasible dimensionality for PCA.
- Apply the fitted sequence of steps to all samples
- Scale all features within each plate using the negative controls from the same plate.

We did not use the post-processing steps recommended for the CPJUMP1 dataset, as it is specifically tailored to CellProfiler-derived features. Instead, we adapted the OpenPhenom recommended post-processing steps to account for the limited number of negative controls in our dataset, which restricts the maximum dimensionality of the PCA step.

Note that while the phenotypic activity benchmark offers an interpretable and practical framework for evaluating DL models in phenotypic drug discovery, performance is highly sensitive to the design of the post-processing pipeline. For example, pipelines optimized for CellProfiler features may not generalize well to DL-derived features, and the effectiveness of each step often depends on the availability and quality of metadata. Even seemingly minor choices, such as aggregating features using the mean versus the median (Table 2), can shift the relative performance of models. These findings highlight that pipeline components should not be treated as modular or interchangeable. Each step must be carefully designed in the context of the full workflow, especially when generative models are involved.

A.5 Additional Results

In Figure 2 we show the results for MAE and EMD across the LSUN and CPJUMP1 datasets, illustrating how these metrics are similar in range for both types of datasets. LSUN images were used to establish a baseline for these metrics, since the Church and Classroom subsets we looked at are natural images similar to the images used to train the SD-VAE.

In Figure 3 we show instead the FID and KLD cross the LSUN and CPJUMP1 subsets, illustrating how the Cell Painting latents appear to be slightly less regularized than the latent embeddings of the natural images following application of the SD-VAE encoder.

Note that the results may be affected by how we chose to group the channels for the InceptionV3 model, which requires 3-channel inputs. This is particularly relevant given the noticeable variation in FID scores across channels. At the 24-hour time point, some channels show nearly twice the FID values compared to others (see Figure 3). We leave examining the effects of channel grouping on the metrics to future work.

Interestingly, we observed that the OpenPhenom features benefit more from mean aggregation, while InceptionV3 features consistently perform better with median aggregation (Table 2); this may reflect further differences in how the two models handle outliers or noise in morphological profiles.

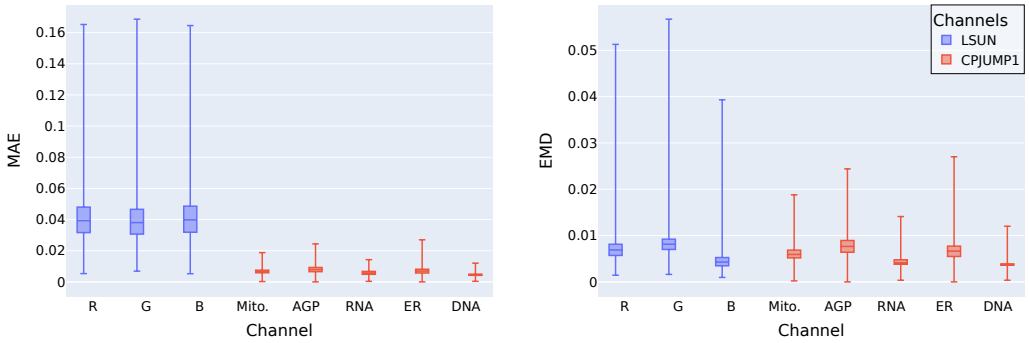

Figure 2: Box plots showing MAE (left) and EMD (right) values across the LSUN and CPJUMP1 datasets and their channels, computed after a single run of SD-VAE applied to all images. The central line within each box shows the median, while the box boundaries represent the 1st and 3rd quartiles.

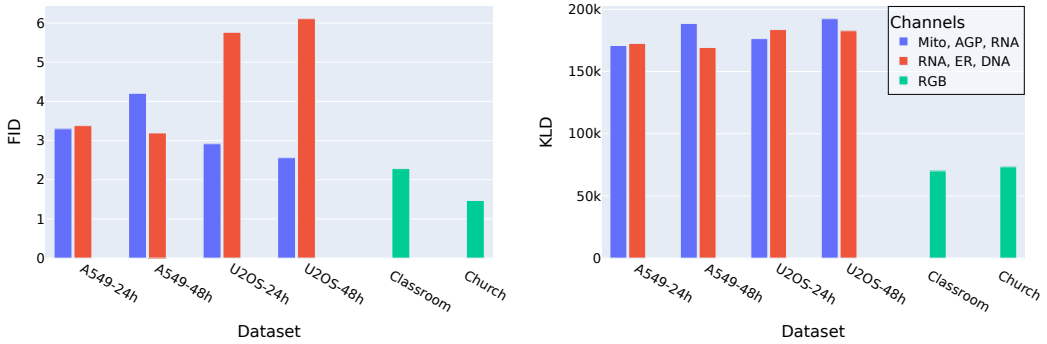

Figure 3: FID (left) and KLD (right) scores across the various data subsets, computed after a single run of SD-VAE applied to all images. FID scores are computed using real and reconstructed images. All images were featurized using InceptionV3. KLD scores are presented as the mean values computed across all samples within each dataset. Standard deviations are approximately 19k, 8k, and 10k for Cell Painting, Classroom, and Church images, respectively.

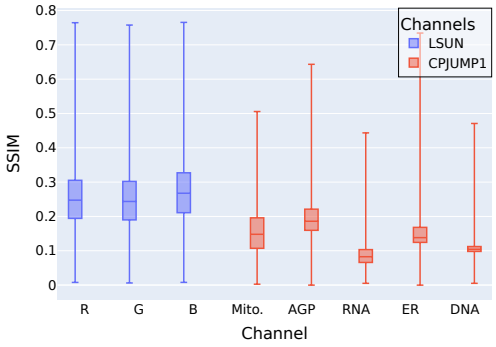

Figure 4: Box plot showing SSIM values across the LSUN and CPJUMP1 datasets and their respective channels, computed after a single run of SD-VAE applied to all images. The central line within each box shows the median, while the box boundaries represent the 1st and 3rd quartiles.

Regardless of the aggregation method, InceptionV3 features outperformed OpenPhenom accross all evaluated conditions.

Table 2: Fraction retrieved (FR) across cell types and time points using different aggregation strategies (mean plate scaling of DL features). All experiments using random features resulted in a FR of 0. The reported values are averaged over three independent runs of the complete pipeline (SD-VAE and information retrieval). For experiments using mean aggregation, the maximum standard deviation is $\sigma = 0.006$, and for median aggregation $\sigma = 0.008$. Experiments conducted on original images by design show no variation ($\sigma = 0$). Best value for each cell line and time point is shown in bold.

			A549		U2OS	
Features	**Data**	**Aggregation**	**24h**	**48h**	**24h**	**48h**
OpenPhenom	Original	Mean	0.804	0.925	0.774	0.719
OpenPhenom	SD-VAE	Mean	0.821	0.915	0.827	0.768
InceptionV3	Original	Mean	0.846	0.935	0.833	0.768
InceptionV3	SD-VAE	Mean	0.846	0.915	**0.852**	0.750
OpenPhenom	Original	Median	0.722	0.882	0.817	0.660
OpenPhenom	SD-VAE	Median	0.729	0.879	0.836	0.697
InceptionV3	Original	Median	0.873	**0.961**	0.837	**0.837**
InceptionV3	SD-VAE	Median	**0.906**	0.951	0.847	**0.837**
CellProfiler [4]	Original	Median	0.761	0.954	0.775	0.663

We noticed that achieving high FR is more difficult for the U2OS cell line than for A549, even though U2OS was included in the OpenPhenom training set. The difference for this discrepancy remains unclear, but may be due to greater phenotypic heterogeneity in the U2OS cell line.

A.6 Hardware and Compute Resources

To facilitate parallelization across multiple GPUs, the datasets are divided into several subsets. To avoid the overhead of saving latent and reconstructed images, the MAE, SSIM, EMD, and KLD metrics are computed and saved in real time during inference in a batch setting. Features extracted from InceptionV3 and OpenPhenom are also saved. Inferencing the datasets was completed in just a few hours by leveraging dozens of NVIDIA A40 GPUs in parallel (up to 340 GPUs). On our facilities, running the whole pipeline, from downloading the images, preprocessing them, featurizing them, all the way to the post-processing and final analysis takes around 8 hours, if exploiting the parallelism of multiple jobs at the same time.