PhenDiff: Revealing Subtle Phenotypes with Diffusion Models in Real Images

Anis Bourou[1,2*], Thomas Boyer[1*], Marzieh Gheisari[1], Kévin Daupin[2], Véronique Dubreuil[2], Aurélie De Thonel[2], Valérie Mezger[2], and Auguste Genovesio[1✉]

ENS[1], Université de Paris Cité[2]

Abstract. For the past few years, deep generative models have increasingly been used in biological research for a variety of tasks. Recently, they have proven to be valuable for uncovering subtle cell phenotypic differences that are not directly discernible to the human eye. However, current methods employed to achieve this goal mainly rely on Generative Adversarial Networks (GANs). While effective, GANs encompass issues such as training instability and mode collapse, and they do not accurately map images back to the model's latent space, which is necessary to synthesize, manipulate, and thus interpret outputs based on real images. In this work, we introduce PhenDiff: a multi-class conditional method leveraging Diffusion Models (DMs) designed to identify shifts in cellular phenotypes by translating a real image from one condition to another. We qualitatively and quantitatively validate this method on cases where the phenotypic changes are visible or invisible, such as in low concentrations of drug treatments. Overall, PhenDiff represents a valuable tool for identifying cellular variations in real microscopy images. We anticipate that it could facilitate the understanding of diseases and advance drug discovery through the identification of novel biomarkers.

Keywords: generative models · subtle phenotypes · microscopy images.

1 Introduction

The emergence of automated imaging and high-throughput platforms has made image-based cellular profiling essential for identifying phenotypic variations [1,2]. Traditional methods such as cell segmentation and quantification of characteristics such as intensity, shape, and texture, are commonly used to identify cellular changes in microscopy images [3]. However, these feature shifts are often challenging to interpret. This difficulty is compounded by the inherent variability among cells within the same condition, which can mask the differences between distinct conditions and complicate accurate analyses [4]. Detecting subtle visual

✉ Correspondence: auguste.genovesio@ens.psl.eu

* Equal contribution

differences between cells in biological images poses a significant challenge but also offers substantial potential for advancing disease understanding, discovering new biomarkers, and developing drugs and diagnostics [5,6]. Recently, generative models have been explored to uncover and analyze cellular phenotypes in microscopy images [4,7,8].

In [8], the approach to identify cellular variations was framed as an image-to-image translation task between 2 classes, similar to methodologies found in [9,10]. The core concept of this method involves training GANs to translate images between two conditions. However, typical High Content Screening (HCS) campaign test for a vast range of conditions, such as different concentrations of potential drugs. This complexity makes the method described in [8] less practical for scenarios involving more than 2 conditions.

In Phenexplain [4], the authors proposed training a conditional StyleGAN [11] to identify cellular changes by performing interpolation in the latent space. Unlike the approach in [8], this method accommodates training across multiple conditions. However, a significant limitation is that cellular changes are identified on *synthetic* images rather than real ones, which may limit the method's applicability. A similar approach was presented in [7]; however, instead of leveraging the latent space of GANs, the authors proposed learning a representation space using self-supervised learning techniques [12].

These methods commonly employ GANs, which are known to suffer from limitations such as training instability and mode collapse [13]. Recently, Diffusion Models (DMs) [14,15,16] have emerged as the new standard in the field of generative models, successfully addressing many of the challenges associated with GANs.

In this work we introduce **PhenDiff**: a novel approach utilizing *multi-class* conditional DMs to translate *real* cell images to other conditions, allowing to spot subtle phenotypic differences triggered by a perturbation. Our code is openly available **on GitHub**.

2 Methods

PhenDiff is built on *Denoising Diffusion Implicit Models* (DDIMs) [14]. It comprises two stages: image inversion and image generation, as shown in Fig. 1. A similar approach was proposed in DDIBs [17] where the authors proposed an image-to-image translation method that relies on two DMs trained independently on each domain. In our approach, we train a *single*, *conditional* DM on multiple domains simultaneously. In this section we first provide an overview of DMs and then dive in the details of our approach.

2.1 Background

Denoising Diffusion Probabilistic Models (DDPMs) are one of the earliest and most successful DMs. They are latent variable models that make use of two

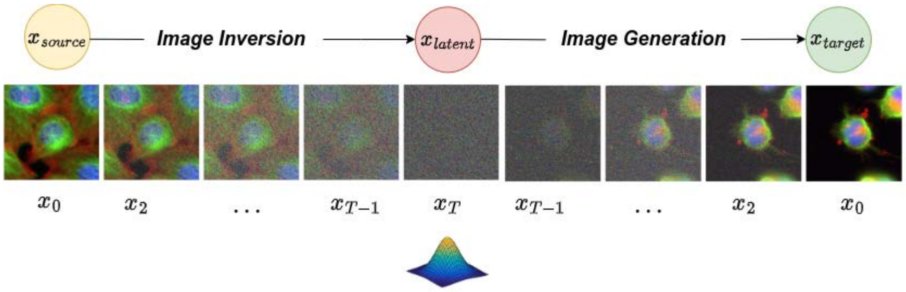


Fig. 1: On this example of the BBBC021 dataset, we translated an untreated image to a treated image with the highest concentration of Latruculin B, notice how the model is able to reproduce the phenotype of the target condition (lower cell count due to the toxicity, no actin cytoskeleton (red channel)) on the real image.

Markovian processes: a fixed forward process that turns data into noise and a learned backward process that turns noise back into the data distribution.
Formally, given a data distribution $x_0 \sim q(x_0)$, the forward process iteratively perturbs the data by adding Gaussian noise to it at successive timestamps. When conditioned on x_0, the joint distribution of the noised images at timestamps $1, 2, ..., T$, denoted as $x_1, x_2, ..., x_T$, can be factorized as follows:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

with the transition kernel $q(x_t|x_{t-1})$ given as:

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

where (β_t) are fixed hyperparmaters. In the backward process the noise is gradually removed by using a learnable transition kernel given by:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

Similar to latent variable models, DMs can be trained using the Variational Lower Bound. In DDPMs [15], the authors derived the following simplified objective function to minimize:

$$\mathbb{E}_{x_0 \sim q(x_0), t, x_t \sim q(x_{1:t}|x_0)}[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2] \tag{4}$$

where ϵ_θ is a learnt function that predicts the noise $\epsilon \sim N(\mathbf{0}, \mathbf{I})$ added to x_t by the forward process, ϵ_θ is often parmeterized by a UNet [18] network. DDPMs require many iterations at inference time to generate satisfying images. To speed up the inference, Denoising Diffusion Implicit Models (DDIMs) [14] can be utilized. Importantly, DDIMs also enjoy another compelling property: deterministic sampling. This allows *exact* inversion, which is instrumental in our approach where we aim at observing phenotypic changes in real images.

2.2 Deterministic Conditional Image Generation

Sampling from a DM corresponds to gradually removing noise from noised images. As described in [19,14], when using the DDIMs deterministic formulation, given x_t, a noised image at timestamp t, the denoised version of it at the timestamp $t-1$ is given by the following formula:

$$x_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} x_t + \sqrt{\alpha_{t-1}} \gamma_t \, \epsilon_\theta(x_t, t, y) \tag{5}$$

where $\gamma_t = \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right)$, $\alpha_t = \prod_{i=1}^{t} (1 - \beta_i)$, and $\epsilon_\theta(x_t, t, y)$ is the predicted noise. We repeat this operation starting from x_T, which corresponds to pure Gaussian noise, to x_0, which is the generated image. Conditional generation is achieved by giving the class label y as additional input to ϵ_θ.

2.3 Image Inversion

Image inversion is the task of finding a latent code that generates back a given real image. It plays a major role in image editing models [17,19,20]. GANs inversion methods are based either on optimization or on learning an image-to-latent encoder [21]. Despite recent progress, GAN inversion remains challenging due to the reduced dimensionality of the latent space in comparison to the image pixel space, as opposed to DMs. In DDIMs, a unique inverted latent code can be obtained [19,16], without any additional optimization or encoding network. In the limit of small steps, The inversion formula is as follows:

$$x_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} x_t + \sqrt{\alpha_{t+1}} \bar{\gamma}_t \epsilon_\theta(x_t, t, y) \tag{6}$$

where $\bar{\gamma}_t = \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right)$.

3 Experiments and Results

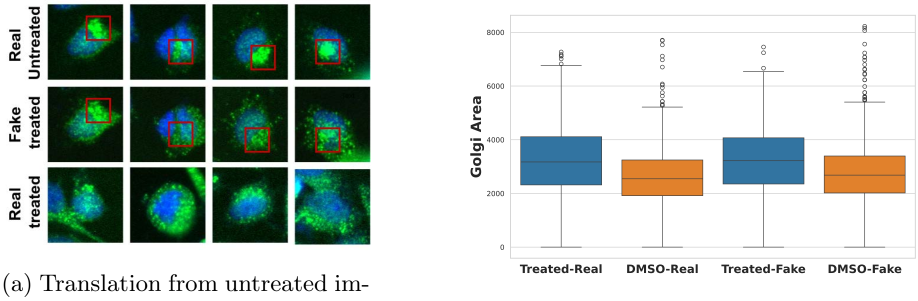
In this section, we first present the datasets we used in our evaluation. We assess cellular variations across the conditions of these datasets both qualitatively and quantitatively. Subsequently, we compare our approach with those based on GANs. Our models were trained using 3 V100 GPUs. The network architecture is a U-Net with 3 ResNet blocks per encoder/decoder and approximatively 70M parameters.

3.1 Datasets

BBBC021: The BBBC021 [22] is a publicly available dataset containing images obtained with fluorescent microscopy of MCF-7 breast cancer cells treated with 113 small molecules across eight concentrations. In our research, we specifically used images of untreated cells and cells treated with 8 concentrations of of 3 compounds: Latrunculin B, Nocodazole, and Herbimycin A (25 conditions altogether). In Fig. 3, the green, blue and red channels label for B-tubulin, DNA and F-actin respectively.

Golgi: Fluorescent microscopy images of HeLa cells untreated (DMSO) and treated with Nocodazole. In Fig. 2a, the green and blue channels label for B-tubulin and DNA respectively.

3.2 Reliable synthesis of visible cell phenotypes



(a) Translation from untreated images to treated ones.

(b) Boxplot of the Golgi area (in pixels).

Fig. 2: We translated real untreated images from the Golgi dataset to treated ones, we can see that Phendiff is able to replicate the effect of the treatment. Left: the Golgi apparatus (red box) is compact in untreated images and splitted in the treated ones (real and generated). Right: An image analysis measurement (Golgi apparatus area) performed on real and synthetic images of both conditions led to the same quantitative conclusion. A two-sided t-test yields a p-value of 1.1e-28 for real images and a p-value of 1.1e-14 for generated images.

In Fig 2a we translate real images of untreated cell to treated images with PhenDiff. We can observe changes in the true morphology of the Golgi apparatus following treatment with Nocodazole: the apparatus has fragmented into smaller stacks. PhenDiff is able to generate translated images that match this phenotype. To validate this observation quantitatively, we computed the area (measured in pixels) occupied by the the Golgi apparatus (green channel). In Fig 2b we can see that there is a significant difference in the mean areas between untreated and treated cells in real images. The same difference is observed in the generated images, which indicates that our method is able to replicate the effects of this treatment and would lead to the same conclusion (see Appendix B for more examples) .

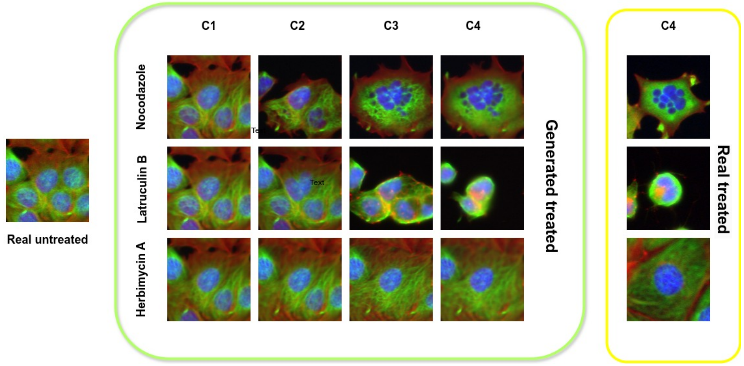


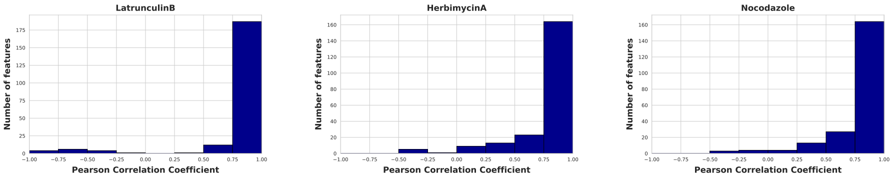
Fig. 3: Translating a single real untreated image into treated counterparts for a given treatment (row) with increasing concentrations (columns, C1, C2, C3, C4). All images within the green rectangle are generated. Within the yellow rectangle are real images treated with corresponding drugs at concentration C4. The true changes induced by the highest concentration closely resemble those in the generated images at C4.

3.3 Reliable synthesis of invisible cell phenotypes

In the previous section, we demonstrated that the obvious phenotypic changes triggered by a treatment could be reliably reproduced. To prove that this method is also valid for detecting invisible phenotypic variations, we translated untreated real images into different classes, each corresponding to a treatment with a given concentration: [C1, C2, C3, C4]; specifically, the concentrations in μM are the following: for Nocodazole $[0.003, 1.0, 3.0, 30.0]$, for Latruculin B $[0.003, 1.0, 3.0, 30.0]$ and for Herbimycin $[0.003, 0.3, 1.0, 10.0]$. We extracted 215 features from 1,000 translated (to all the conditions) and real images using CellProfiler [3]. For each feature, we calculated the mean values for both the real and generated images for each condition (a condition being a treatement at one concentration). We then computed the correlation between the real and synthetic mean values for that specific feature across all concentrations of a given treatment (see Appendix A). Fig. 4(a)-(c) display histograms of the correlation values, showing that, for the large majority of CellProfiler features, there is a strong correlation between the generated and real images. This indicates that our method can faithfully reproduce cellular variations across all concentrations including the lowest ones, for different treatments, demonstrating that synthetic images displaying invisible phenotypes can also be accurately reproduced.

3.4 Subtle phenotypic variation can be identified on low concentrations

Fig. 3 shows that the generated images of low concentrations of distinct treatments are slightly different. As we increase the concentrations, some cells at the border of the images are systematically eliminated due to the toxicity of the treatments, although not in the same way and at the same concentrations. Moreover, Latrunculin B tends to contract the cytoplasm, whereas Nocodazole tends to extend it with increasing concentrations, something hardly visible in real images. At the highest concentration, the phenotype changes induced by the treatments closely resemble those observed in the real images.



(a) Latrunculin B (b) Herbimycin A (c) Nocodazole

Fig. 4: Distribution of Pearson correlations coefficients between the features extracted in real and translated images for each treatment. We observe that the majority of features in the real and the translated images are strongly correlated which indicates that the proposed method is able to recover the cellular variations in the microscopy images.

Table 1: The FID scores for the translated images to all the concentrations of Latrunculin B treatment, the lower the score the better.

Method	C1	C2	C3	C4	C5	C6	C7	C8
StarGAN	91.17	86.57	89.39	86.57	108.53	114.89	119.47	111.72
CycleGAN	31.59	30.7	28.4	34.18	32.85	21.53	20.38	24.57
PhenEx/W	25.07	24.6	22.31	28.93	40.6	91.86	131.49	137.95
PhenEx/W+	36.48	36.62	31.29	42.80	58.17	148.32	203.15	220.63
PhenDiff (ours)	**22.19**	**21.18**	**22.19**	**21.41**	**23.37**	**18.38**	**14.26**	**16.6**

3.5 Evaluation of PhenDiff against existing methods

We recall that Phenexplain [4], a method leveraging conditional StyleGAN, was introduced for identifying subtle phenotypic variations [4]. However, its application to real images is limited. To address this, we adapt it for use with real images. This adaptation involves mapping real images into the latent space of a pre-trained StyleGAN model. We identify the vector representing the transition from the image's current class to a target class by calculating the difference between the average latent codes for each class. This allows us to move the image's latent code along this vector to produce the translated image.

Adapting Phenexplain for real images required StyleGAN inversion, a step marked by challenges. The inversion quality, essential for our task, varies depending on the use of the original W or extended W^+ latent spaces, each offering different trade-offs between reconstruction fidelity and editability [23,24]. Acknowledging the trade-offs between using the W and W^+ spaces, we developed two versions of the adapted method: PhenEx/W and PhenEx/W+.

Table 2: Comparing the quality of the reconstructed real untreated image using the L2 loss

Method	StarGAN	PhenEx/W	PhenEx/W+	PhenDiff
Reconstruction Loss	1986.80	2001.09	1895.58	**415.15**

Additionally, considering the use of a variant of CycleGAN for the task of identifying subtle phenotypic variation in real images in [8], we also included CycleGAN as another baseline for comparison. However, this model is limited to translations between no more than two classes. To evaluate our method against a model capable of multi-class translations, we included StarGAN [25], a representative method for multi-domain image-to-image translation, into our baseline comparisons. For the evaluations, PhenDiff was trained on all the concentrations of Latrunculin B, a process replicated for StarGAN. Due to its limitation in supporting multi-domain image-to-image translation, CycleGAN necessitated the training of eight different models, each enabling translation between the untreated class and another concentration.

In Table 1, we applied our method to translate 1,000 untreated images to the 8 concentrations of Latrunculin B, evaluating the generated images quality using the Fréchet Inception Distance (FID) score [26]. StarGAN, with its design focus on natural images, exhibits high FID values across all classes, suggesting its lower effectiveness for biological image translation. CycleGAN demonstrates moderate translation quality with acceptable FID scores but requires training eight different models for each treatment, making it computationally intensive. PhenEx/W shows better performance at lower concentrations compared to PhenEx/W+, yet both struggle with accurately replicating effects at higher doses, as indicated by increasing FID scores.

The quality of the reconstructed real untreated images, as shown in Table 2, is crucial for our analysis. Here, PhenDiff stands out by achieving the lowest reconstruction loss, emphasizing its enhanced ability to detect cellular variations in real images (see Appendix C). The development of PhenEx/W and PhenEx/W+ was motivated by the challenges of GAN inversion, seeking to balance reconstruction fidelity and editability. Despite this, the marginal difference in their reconstruction losses suggests a subtle balance in their capabilities. Overall, these results highlight our method's superiority in image generation quality and inversion accuracy compared to the baseline models, demonstrating its effectiveness in handling the complexities of biological image translation.

4 Conclusion

In this work, we introduced PhenDiff, a multi-class image-to-image translation method leveraging conditional diffusion models to identify subtle phenotypic variations in real microscopy images. Our experiments demonstrate that PhenDiff can accurately produce variations in phenotypes induced by various treatments. Additionally, compared to existing methods, particularly those based on Generative Adversarial Networks (GANs), PhenDiff exhibits superior performance in terms of image quality. Moreover, its precise image inversion capability enables the detection of these variations in real images. Overall, our findings suggest that PhenDiff can be a valuable tool in understanding the effects of certain treatments and in identifying new biomarkers.

Acknowledgments. This work was supported by the RUBINeuroStress ANR-19-CE16-0030, the Fondation de la Recherche Médicale (Équipe FRM Equ201903007924). ANR–10–LABX–54 MEMOLIFE, ANR–10 IDEX 0001–02 PSL* Université Paris, ANR du Labex Who Am I? (ANR-11-LABX-0071), Idex ANR-18-IDEX-0001 and was granted access to the HPC resources of IDRIS under the allocation 2020-AD011011495 made by GENCI.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca Senft, Yu Han, Mehrtash Babadi, Peter Horvath, Beth A. Cimini, Anne E. Carpenter, Shantanu Singh, and Juan C. Caicedo. Learning representations for image-based profiling of perturbations. *bioRxiv*, 2022.
2. Srinivas Niranj Chandrasekaran Chandrasekaran, Hugo Ceulemans, Jeffrey D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20:145–159, 2021.
3. Anne Carpenter et al. Cellprofiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7:R100, 02 2006.
4. Alexis Lamiable et al. Revealing invisible cell phenotypes with conditional generative modeling. *Nature Communications*, 14, 2022.
5. Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
6. Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), 2023.
7. Ruben Fonnegra, Mohammad Sanian, Zitong Chen, Lassi Paavolainen, and Juan Caicedo. Analysis of cellular phenotypes with image-based generative models. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

8. Anis Bourou and Auguste Genovesio. Unpaired image-to-image translation with limited data to reveal subtle phenotypes, 2023.
9. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.
10. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.
11. Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.
12. Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends, 2023.
13. Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2020.
14. Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.
15. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.
16. Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.
17. Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *The Eleventh International Conference on Learning Representations*, 2023.
18. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.
19. Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.
20. Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation, 2022.
21. Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey, 2022.
22. Peter D. Caie et al. High-Content Phenotypic Profiling of Drug Response Signatures across Distinct Cancer Cells. *Molecular Cancer Therapeutics*, 9(6):1913–1926, 06 2010.
23. Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? pages 4432–4441, 2019.
24. Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in stylegan using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020.
25. Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2018.
26. 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, 2018.