LEA: Latent Eigenvalue Analysis in application to high-throughput phenotypic drug screening

Anonymous authors
Paper under double-blind review

Abstract

Understanding the phenotypic characteristics of cells in culture and detecting perturbations introduced by drug stimulation is of great importance for biomedical research. However, a thorough and comprehensive analysis of phenotypic heterogeneity is challenged by the complex nature of cell-level data. Here, we propose a novel Latent Eigenvalue Analysis (LEA) framework and apply it to high-throughput phenotypic profiling with single-cell and single-organelle granularity. Using the publicly available SARS-CoV-2 datasets stained with the multiplexed fluorescent cell-painting protocol, we demonstrate the power of the LEA approach in the investigation of phenotypic changes induced by more than 1800 drug compounds. As a result, LEA achieves a robust quantification of phenotypic changes introduced by drug treatment. Moreover, this quantification can be biologically supported by simulating clearly observable phenotypic transitions in a broad spectrum of use cases. In conclusion, LEA represents a new and broadly applicable approach for quantitative and interpretable analysis in routine drug screening practice.

1 Introduction

In the cell-based drug screening, the emergence of novel fluorescent imaging protocols allows to reveal relevant cellular components and organelles (*e.g*., Nucleus (DNA), Endoplasmic reticulum (ER), Actin (Actin), Nucleolus and cytoplasmic RNA (RNA), Golgi and plasma membrane (Golgi)) in a highly multiplexed, high content manner. Following the cell-painting protocol (Bray et al., 2016), two large-scale drug screening datasets RxRx19 (a, b) (Cuccarese et al., 2020) have been recently released, which include more than 1800 drug compounds with up to 8 different concentrations that are tested on 3 different cell-lines. These high-throughput phenomic libraries have indeed stimulated the development of novel approaches for analyzing phenotypic effects introduced by drug treatments (Cuccarese et al., 2020). However, current methods failed to fully utilize the single-cell and highly multiplexed nature of these datasets, leaving much to be discovered.

In existing studies (Cuccarese et al., 2020; Koh et al., 2021), researchers usually start the analysis by downsizing the raw image read-outs or by exclusion of some fluorescent channels (Koh et al., 2021). In a supervised manner, classification features of the entire image can then be learned to determine the phenotypic changes of a cell population induced by different drug compounds and can be linked to treatment efficacy (Cuccarese et al., 2020) (Fig. 1 (a)). But, it is still an open and essential question to conduct a more in-depth analysis at the single-cell and/or single-organelle level for understanding the drug effects in a concentration-specific manner. From a technical perspective, handling such massive datasets poses further statistical challenges (Johnstone & Titterington, 2009). New approaches for the comprehensive quantification of data heterogeneity are needed for the analysis of high-dimensional datasets (Wiles et al., 2021) in different domains.

In the biomedical domain, statistical tests such as the F-test (Richard & Hahs-Vaughn, 2007) and Student's t-test (Owen, 1965) are commonly used to examine the statistical discrepancy between two collections of heterogeneous tabular data. Despite demonstrable success in theoretical studies (Peters et al., 2016; Heinze-Deml et al., 2018; Gamella & Heinze-Deml, 2020), it is far from straightforward to apply them to real-world multi-dimensional cases. For instance, Wu et al. (2022) found that the p-values computed with these statistical tests (Richard & Hahs-Vaughn, 2007; Owen, 1965; Levene, 1961; Wilcoxon, 1992) are not robust when applied to real-world clinical data, and then undermine the accuracy of identifying causal associations between diagnostic features and patient outcome.

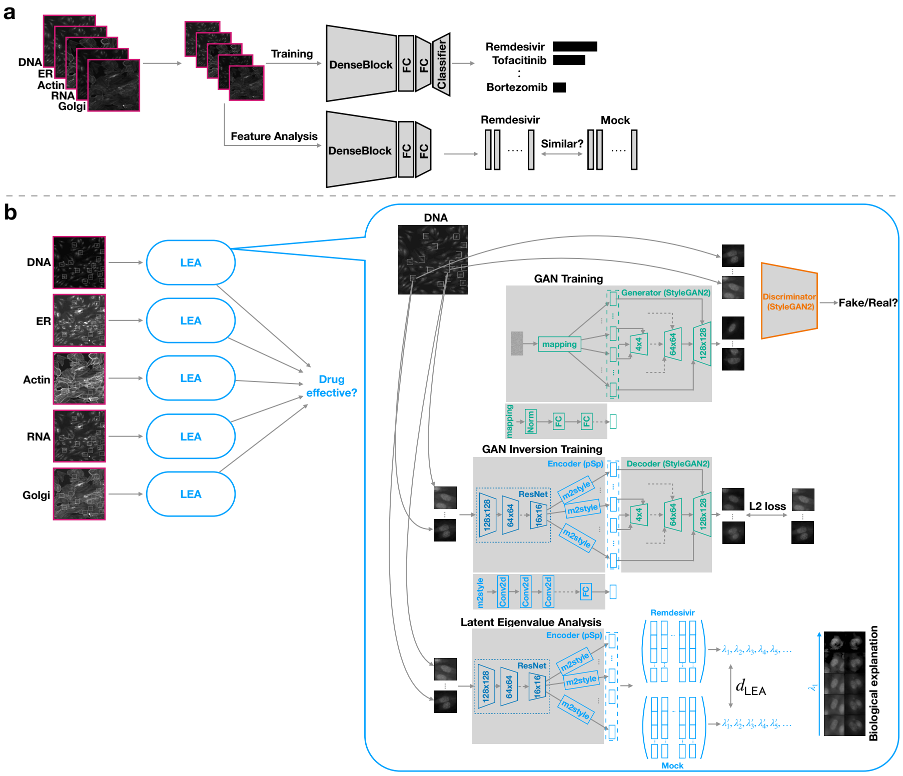


Figure 1: **Model illustrations for the baseline method (Cuccarese et al., 2020) (a) and proposed single-cell LEA approach (b)**. **a**, In the baseline method, a variant model of DenseNet-161 is trained on the downsized multiplexed fluorescent images for classifying the drug compounds (Remdesivir, Tofacitinib, Bortezomib shown as exemplars). Then, the learned features are utilized to analyze the effectiveness of different drugs. **b**, In LEA, we start the pipeline by pre-training the decoder on center-cropped single-cell images in an unsupervised manner for each fluorescent channel. Then, we learn robust latent representations with a residual-based encoder (Alaluf et al., 2021) for reconstructing these single-cell images. For quantifying the drug effects, we compute the eigenvalues with learned single-cell representations and support our quantitative results with clearly observable phenotypic transitions.

In the deep learning domain, the importance of measuring the difference (heterogeneity) between fake and real data has been recognized in parallel to the development of generative adversarial nets (GAN) (Goodfellow et al., 2014; Heusel et al., 2017; Karras et al., 2020). To measure the quality of GAN reconstructions, researchers have proposed a variety of evaluation methods such as Fréchet Inception Distance (d_{FID}) (Heusel et al., 2017), Inception Score (Salimans et al., 2016) and Kernel Inception Distance (d_{KID}) (Bińkowski et al., 2018). These approaches could be used to differentiate image data distributions for the analysis of biomedical datasets. Nevertheless, it is not trivial to derive a multi-dimensional quantitative understanding with these scores, nor can we directly support them with plausible visual explanations. Therefore, they are less satisfactory for critical biomedical applications. Motivated by the emergence of (unsorted) eigenvalues in the improved implementation of d_{FID}, Wu & Koelzer (2022) recently suggested comparing sorted eigenvalues (d_{Eig}) as a simple alternative to d_{FID}. For $i = 1, 2$, let $\mathbf{Z}_i := (\mathbf{z}^i_1, \ldots, \mathbf{z}^i_{n_i})$ be a collection of n_i p-dimensional vectors. This leads to the following definition that can provide informative measurements along principal axes and facilitate a more complete analysis of data heterogeneity:

Definition 1. *Let* $\mathbf{S}_i = \frac{1}{n_i}\mathbf{Z}_i\mathbf{Z}_i^\top$ *be the sample covariance matrix (SCM) of* $\mathbf{Z}_i$*, then we define*

$$d_{\mathsf{Eig}}(\mathbf{S}_1, \mathbf{S}_2)^2 = \sum_{j=1}^{p} (\sqrt{\lambda_1^j} - \sqrt{\lambda_2^j})^2, \tag{1}$$

where λ_i^j *is the* j*-th largest eigenvalue of* $\mathbf{S}_i$.

2 Proposed LEA

Quantification of phenotypic heterogeneity. Based on the theoretical foundation behind d_{Eig} (Wu & Koelzer, 2022), we propose a novel latent eigenvalue analysis (LEA) for high-throughput phenotypic profiling (Fig. 1). In the study of d_{Eig}, $\mathbf{Z}_i$ is usually the collection of features obtained with the penultimate layer (*pool3*) of an Inception V3 model (Szegedy et al., 2016), where the model is trained for an ImageNet classification task. However, such an Inception model trained with ImageNet is not suitable for deriving meaningful features of multiplexed single-cell images. Alternatively, we utilize the approach of GAN inversion (Xia et al., 2022) and propose to learn the latent representations $\mathbf{Z}_{i,c}$ on the c-th fluorescent channel of center-cropped single-cell images (Fig. 1 (b)). To address the infeasible deployment of d_{Eig} to cases where we obtain imbalanced values of d_{Eig} *w.r.t.* different channels, we subsequently propose

Definition 2. *For $i = 1, 2$ and $k = 1, \dots, c$, let $\mathbf{S}_i = (\frac{1}{n_{i,1}}\mathbf{Z}_{i,1}\mathbf{Z}_{i,1}^\mathsf{T}, \dots, \frac{1}{n_{i,c}}\mathbf{Z}_{i,c}\mathbf{Z}_{i,c}^\mathsf{T})$ be the collection of SCMs of $\mathbf{Z}_{i,k}$, then we define*

$$d_{\mathsf{LEA}}(\mathbf{S}_1, \mathbf{S}_2) = \sum_{k=1}^{c} \sum_{j=1}^{p'} \frac{(\sqrt{\lambda_{1,k}^j} - \sqrt{\lambda_{2,k}^j})^2}{\lambda_{1,k}^j}, \tag{2}$$

where $p' \ll p$ and $\mathbf{S}_1$ is the reference SCM.

Similar to Principal Component Analysis (PCA) (Shlens, 2014), we only utilize the $p' \ll p$ largest eigenvalues that reflect the largest variances and the most critical information. As the 5 largest eigenvalues dominate $> 95\%$ of the overall values in the experiments, we set $p' = 5$ throughout the article (Please see also Appendix Sec. A.1 and A.2). The reference $\mathbf{S}_1$ can be concretely determined in a given dataset, *e.g.*, the SCM of mock cell read-outs in this study. We thus propose a novel quantitative method for phenotypic profiling of cells at single-organelle resolution.

Visualization of phenotypic transitions. Complementary to the d_{LEA} that measures the eigenvalue difference along each principal axis, we simulate observable phenotypic transitions by manipulating the principal component(s). This enables direct linkage of the observed phenotypic heterogeneity in the given datasets with human-interpretable biological information. Notably, there have been previous studies in understanding latent semantic transitions for natural images (Shen & Zhou, 2021; Härkönen et al., 2020; Patashnik et al., 2021; Wu et al., 2021). To probe the latent semantics of generative models, many of these investigations conducted image manipulations on fake images, where the manipulations are either unrelated or loosely related to a quantitative measurement. For example, Härkönen et al. (2020) proposed to edit fake images by adding weighted eigenvectors to its latent representations. Similarly, Shen & Zhou (2021) utilized the closed-form factorization to determine the manipulation direction. Since such manipulation directions are not parallel to the principal axes, they cannot be directly used to explain the eigenvalue heterogeneity embedded in d_{LEA}. To resolve this gap, we propose to output a single-cell image sequence by manipulating the largest principal component(s) of latent representations $\mathbf{Z}_{i,k}$ (Def. 2), where $\mathbf{Z}_{i,k}$ are derived from real single-cell images.

Definition 3. *Following the specifications of $\lambda_{i,k}^j, \mathbf{Z}_{i,k}$ of Def. 2, let $\boldsymbol{\lambda}_{i,k} = \mathsf{diag}(\lambda_{i,k}^1, \dots, \lambda_{i,k}^p)$ be decomposed as $\boldsymbol{\lambda}_{i,k} = \frac{1}{n_{i,k}}\tilde{\mathbf{Z}}_{i,k}\tilde{\mathbf{Z}}_{i,k}^\mathsf{T}$, where $\tilde{\mathbf{Z}}_{i,k} = \mathbf{O}_{i,k}^\mathsf{T}\mathbf{Z}_{i,k}$ are the principal components w.r.t. the orthogonal eigenbasis $\mathbf{O}_{i,k}$. Given $j \in \{1, \dots, p'\}$, we interpret the difference between $\lambda_{1,k}^j$ and $\lambda_{2,k}^j$ (Eq. 2) by visualizing $\beta_{i,k}^j\tilde{\mathbf{Z}}_{i,k}$ manipulated on the j-th principal component that brings*

$$\tilde{\boldsymbol{\lambda}}_{i,k}^j = \frac{1}{n_{i,k}}(\beta_{i,k}^j)^2\tilde{\mathbf{Z}}_{i,k}\tilde{\mathbf{Z}}_{i,k}^\mathsf{T} = (\beta_{i,k}^j)^2\boldsymbol{\lambda}_{i,k}, where\ \beta_{i,k}^j = \mathsf{diag}(1, \dots, 1, \beta_{i,k}^j, 1, \dots, 1). \tag{3}$$

To guarantee the consistency and to simulate clearly interpretable phenotypic transitions, image sequences *w.r.t.* $\beta_{i,k}^j\tilde{\mathbf{Z}}_{i,k}$ are always obtained by step-wise assigning $\beta_{i,k}^j = 2^m$ in this study, where $m = (-1.8, -1.2, -0.6, 0, 0.6, 1.2, 1.8)$. By combining the quantification and visualization components, we proposed the Latent Eigenvalue Analysis (LEA) pipeline. Our contribution is two-fold:

- By comparing the largest eigenvalues, we propose the numerically robust quantification d_{LEA} of phenotypic heterogeneity in multiplexed fluorescent image datasets. As a direct application of d_{LEA}, we refine the high-throughput cell-based drug analysis to single-cell and single-organelle granularity.
- By manipulating the largest principal components, we provide phenotypically plausible visual explanations to d_{LEA}. In the context of domain knowledge, these transitions can support novel interpretations of drug effects and drug response heterogeneity.

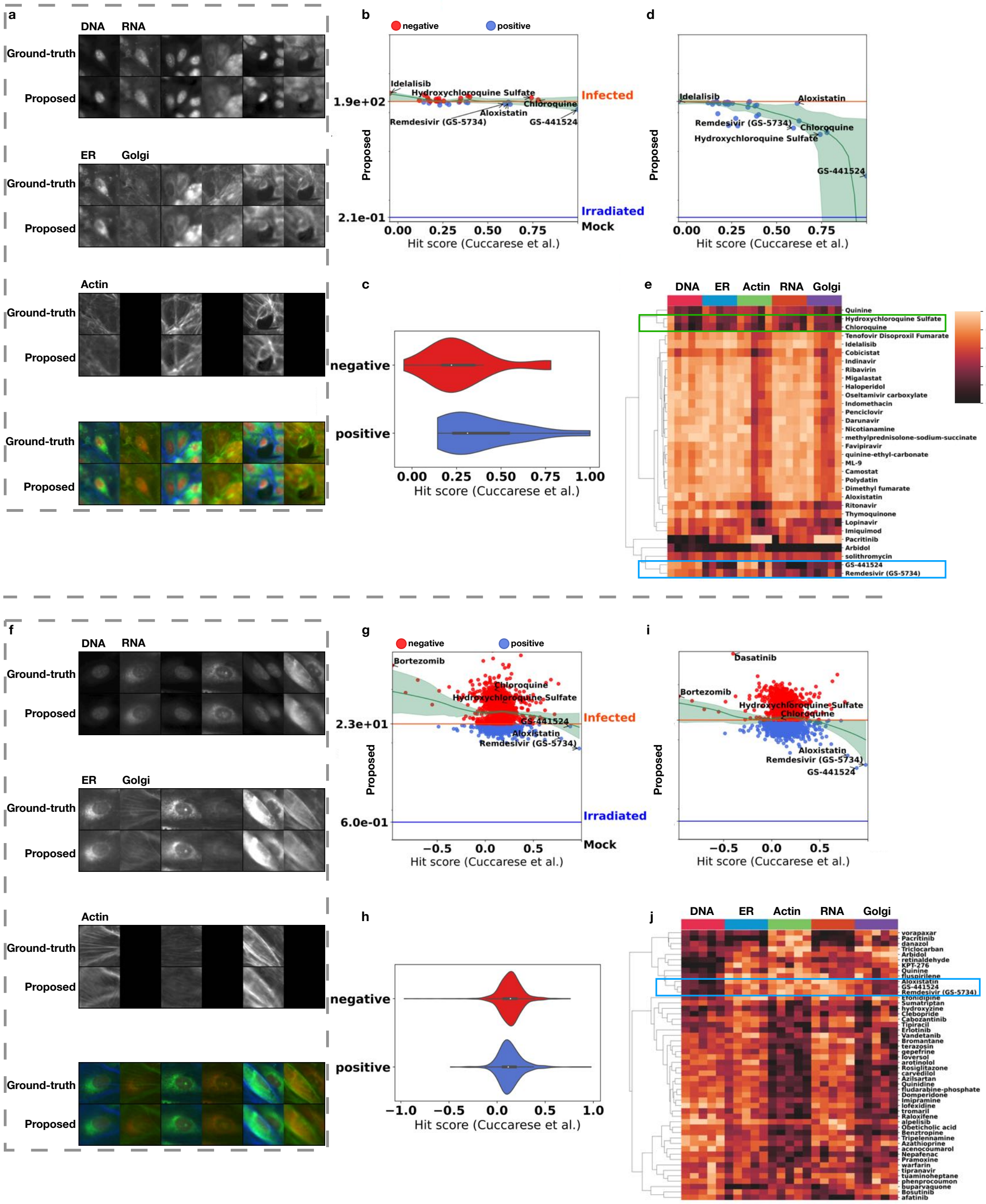


Figure 2: **Reconstruction visualization of LEA and quantitative comparison of drug responses between the baseline (Cuccarese et al., 2020) and d_{LEA} (Proposed).** **a** (VERO) and **f** (HRCE): The reconstructed samples obtained by LEA. **b** (VERO) and **g** (HRCE): The quantitative comparison between the hit score (Cuccarese et al., 2020) and d_{LEA} with the latent representations of all concentrations. **c** (VERO) and **h** (HRCE): The violin plot of overall comparison between the hit score and d_{LEA}. **d** (VERO) and **i** (HRCE): The quantitative comparison between the hit score and d_{LEA} with the latent representations of optimal drug concentration. **e** (VERO) and **j** (HRCE): The hierarchical clustering of top 50 drug compounds (if exist) *w.r.t.* the 5 largest eigenvalues of the latent representations of optimal drug concentration.

3 Results

Here, we report the application of LEA to two large-scale phenomic libraries RxRx19 (a,b) released by Recursion (Cuccarese et al., 2020), which document the effects of more than 1800 drug candidates on Severe Acute Respiratory Syndrome Coronavirus Type 2 (SARS-CoV-2) infection and associated systemic inflammation using the multiplexed fluorescent cell painting protocol on human and animal cell-lines. We set the drug hit score proposed by Cuccarese et al. (2020) as the baseline and demonstrate the performance of d_{LEA}. To investigate the phenotypic effects of drug candidates at single-cell resolution, we carried out cell segmentation using the DNA channel (Please see the Mahotas documentation (Coelho, 2012)). Accordingly, we derive 23 million 64×64 single-cell training images from 0.37 million raw images. This allows us to analyze drug effects on individual cell organelle components, greatly extending the range of detectable phenotypic perturbations. As shown in Fig. 1 (a, b), our approach using unsupervised training on the center-cropped cell images for the individual fluorescent channels indeed differs greatly from the published baseline using supervised training on the entire image read-outs. As the phenomic library (RxRx19a) has four cell conditions: Mock control (Mock), irradiated control (Irradiated), infected without drug treatment (Infected), and infected with different drug treatments (Drug), we set the reference $\mathbf{S}_{1=\mathsf{Mock}}$ (Eq. 2) corresponding to the 'Mock' latent representations and then determine the effect of a drug based on whether it reverses the phenoprint of infected cells. Besides, we report $d_{\mathsf{LEA}} \times 100$ (*e.g.*, percentage) for clearer visualization in the following plots.

Definition 4. *Following the specification of SCMs $\mathbf{S}_{\mathsf{Mock}}, \mathbf{S}_{\mathsf{Infected}}, \mathbf{S}_{\mathsf{Drug}}$ as Eq. 2, we define*

$$\text{A drug is} \begin{cases} \textit{positive}, & \textit{if } d_{\mathsf{LEA}}(\mathbf{S}_{\mathsf{Mock}}, \mathbf{S}_{\mathsf{Drug}}) < d_{\mathsf{LEA}}(\mathbf{S}_{\mathsf{Mock}}, \mathbf{S}_{\mathsf{Infected}}). \\ \textit{negative}, & \textit{else}. \end{cases} \tag{4}$$

LEA benchmarking. Throughout this study, we consider StyleGAN2_pSp the default architecture for conducting drug screening experiments. Concretely, we take StyleGAN2 (Karras et al., 2019) as the decoder and the residual-based 'pixel2style2pixel' (pSp) (Richardson et al., 2021) encoder to build the LEA model. We refer interested readers to Appendix Sec. A.2 for more toy and ablative studies on model designs. For experiments carried out on both human (Human renal cortical epithelial cells, HRCE) and animal (Kidney epithelial cells of African green monkey, VERO) cell-lines, we start our investigation by comparing the reconstruction quality between the proposed LEA – trained on individual fluorescent channels and ensemble LEA – trained on all the channels. For more discussions please see Appendix Sec. A.1.

Overall comparison. For a thorough comparison, we screened all drug compounds tested on VERO and HRCE cell-lines (with the exclusion of three drugs that have duplicated or ambiguous names). Despite fundamentally different model designs (Fig. 1 (a,b)), our quantitative score d_{LEA} demonstrates an overall consistent correlation to the baseline hit score (Cuccarese et al., 2020), that is, the lower the d_{LEA} is, the higher the baseline hit score is. Importantly, the effect estimation for a given drug compound can be directly derived from d_{LEA}, while a manual threshold determination is required for the hit score (Cuccarese et al., 2020). As displayed in Fig. 2 (b, c) of the VERO experiments, d_{LEA} shows a mild yet meaningful decreasing trend with growing hit scores and identifies Remdesivir and its prodrug GS-441524 as efficacious compounds when using all their latent representations that are independent of concentration. Further, LEA allows to take the optimal drug concentration into consideration, and thus achieves a superior resolution in identifying effective drug candidates as illustrated in Fig. 2 (d). Importantly, Remdesivir and GS-441524 remain the top effective candidates among all the drug compounds. The superior effectiveness of Remdesivir and GS-441524 can also be differentiated from other drugs by examining the hierarchical clustering (Fig. 2 (e)). For example, we observe distinct patterns of latent representations of ER, RNA and Golgi channels for Remdesivir and GS-441524. This corresponds to the successfully reversed phenoprint reflected by the largest eigenvalues, *e.g.*, Mock:1.33, GS-441524:1.33, Remdesivir:1.36, and Infected:1.66 for ER ($\times 10^5$); Mock:1.75, GS-441524:1.76, Remdesivir:1.79, and Infected:2.20 for RNA ($\times 10^5$); Mock:3.23, GS-441524:3.20, Remdesivir:3.23, and Infected:3.69 for Golgi ($\times 10^5$). Similar to the VERO experiment, Fig. 2 (g) and (h) show that d_{LEA} remain well correlated with the baseline in the HRCE experiment. Strikingly, Remdesivir and GS-441524 are identified as strongly efficacious when computing the eigenvalues on the latent representations of the optimal drug concentration (Fig. 2 (i)), indicating verifiable positive drug effects achieved by both candidates. On the other hand, Chloroquine and Hydroxychloroquine demonstrate contradictory effects on both the latent representations of the cells treated with different drug concentrations and the optimal drug concentration, both of which are identified as negative by the $d_{\mathsf{LEA}}(\mathbf{S}_{\mathsf{Mock}}, \mathbf{S}_{\mathsf{Infected}})$ threshold (Fig. 2 (g) and (i)). Such inconsistency between the ineffective identification on the human cell-line and the effective identification on the animal cell-line undermines its fidelity in clinical treatment, which can be explained by the fact that neither of them is recommended in treating hospitalized COVID-19 patients according to clinical studies (Roustit et al., 2020; Saghir et al., 2021). If we examine the latent representations in more detail, Fig. 2 (j) highlights the unique patterns presented in ER, Actin, and RNA channels for Remdesivir and GS-441524, which reveal novel and subtle phenotypic changes that were previously unidentified. Cell-level phenotype analysis by d_{LEA} can therefore provide novel insights and patterns in high-throughput drug-screening experiments as candidates for subsequent biological exploration.

Fine-grained quantification and visual interpretation. We report the fine-grained quantification on both VERO and HRCE experiments for individual fluorescent channels and drug concentration levels. As for the overall analysis, the small difference between mock and irradiated control is correctly captured in individual fluorescent channels for both experiments, which serves as an important sanity check for the stratified quantification. Regarding the drugs of interest presented in Fig. 3 (a) and Fig. 4 (a), we observe a consistently improved phenoprint rescue with an increasing dose level of the drug. This reflects the biologically plausible observation of increased inhibitory effects on cellular infection by SARS-CoV-2 with increasing drug concentrations. For VERO and HRCE cell-lines, Remdesivir and GS-441524 are consistently identified as effective compounds *w.r.t.* individual fluorescent channels as well as all channels, which supports the clinical utility of Remdesivir approved by U.S. Food and Drug Administration (FDA) [1]. Meanwhile, the heterogenous effects of Chloroquine and Hydroxychloroquine are also revealed in our refined analysis. For instance, the negative hit results in Actin, RNA and Golgi channels eventually undermine the overall performance of both candidates and provide the negative evidence for both drugs derived from real-world clinical studies (Avezum et al., 2022). Furthermore, the PCA plots displayed in (b) of Fig. 3 and 4 clearly support the efficacious hit results achieved by Remdesivir. With the k-means clustering on the largest principal component(s), we observe meaningful groups based on the nucleus morphology (DNA) for both VERO and HRCE. Besides, interesting and striking phenotypic transitions arise in other understudied channels. Taking the RNA channel as a concrete example, the largest eigenvalue ($\times 10^5$) of mock and infected cells are 1.75, 1.23 versus 2.20, 1.28 for VERO and HRCE *resp.* As shown in (b) of Fig. 3 and 4, the cellular sequences presented from left to right with enlarging the largest principal component(s) imply increased RNA production in the cytoplasm. This observation is biologically plausible, as SARS-CoV-2 expresses RNA-dependent RNA polymerase as well as a large number of supporting factors to transcribe and replicate the viral genome in infected cells (Malone et al., 2022). Viral infection of host cells thus leads to massive upregulation of the production of viral RNA in the cytoplasm, which is correctly identified by the d_{LEA} analysis at subcellular resolution. Importantly, Remdesivir acts as a nucleoside analog and stalls the RNA-dependent RNA polymerase of coronaviruses (Kokic et al., 2021). Our analysis identifies this effect at the phenotypic level, as Remdesivir treatment calms the hyper state reflected by shifting the largest eigenvalue from 2.20 to 1.79 (VERO) and from 1.28 to 1.26 (HRCE). Taking the 5 largest eigenvalues as a whole, we further robustify and differentiate the positive drugs from negative ones, while demonstrating persistent transitions for all channels.

From in vitro to in vivo studies.

In the cell-based in vitro studies, we have shown the refinement and improvement of LEA on the animal (VERO) and human (HRCE) cell-lines. Importantly, the key takeaways of drugs of interest can be well supported by relevant clinical studies. Nevertheless, we acknowledge the limitation of LEA on the human umbilical vein endothelial (HUVEC) cell-line (RxRx19b), which models cytokine storm conditions in severe COVID-19 (Hu et al., 2021). As displayed in Appendix Fig. 9, the overall drug identifications achieved by LEA are less consistent under these conditions with the baseline results. Although the least effective drug candidates identified by (Cuccarese et al., 2020) are likely to be assigned with high d_{LEA} score, the association is less clear in regards to drug candidates with a positive baseline hit score (*e.g.*, c-MET inhibitors in Fig. 9 (b)). Such inconsistency between LEA and the baseline approach (Cuccarese et al., 2020) suggests the current level of evidence remains inconclusive. Whether this observation is due to a high level of variance or a true lack of efficacy of these drug candidates requires follow-up in vivo studies.

[1] `https://www.fda.gov/drugs/emergency-preparedness-drugs/coronavirus-covid-19-drugs`

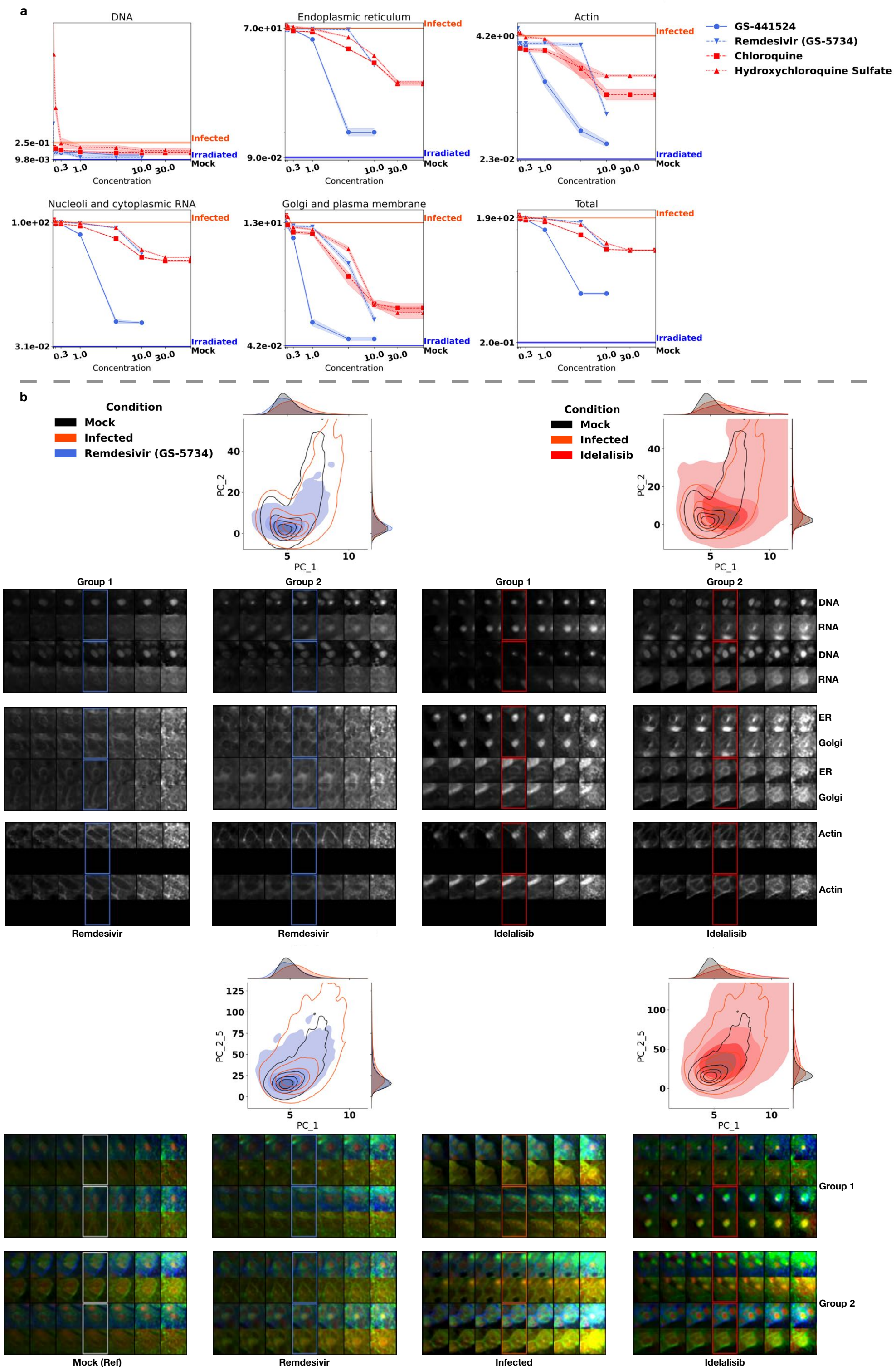


Figure 3: **Identification of drug-concentration dependent effects and visual interpretation for key drugs of interest in the VERO cell-line**. **a**, The proposed d_{LEA} of different drug concentrations for individual and all fluorescent channels. Here, we report the mean d_{LEA} (with standard deviation) averaged on 4 randomly sampled cell collections. **b**, The PCA plots and phenotypic transitions driven by manipulating the largest (top) and 5 largest (bottom) principal component(s). The bounding box indicates the reconstructed image.

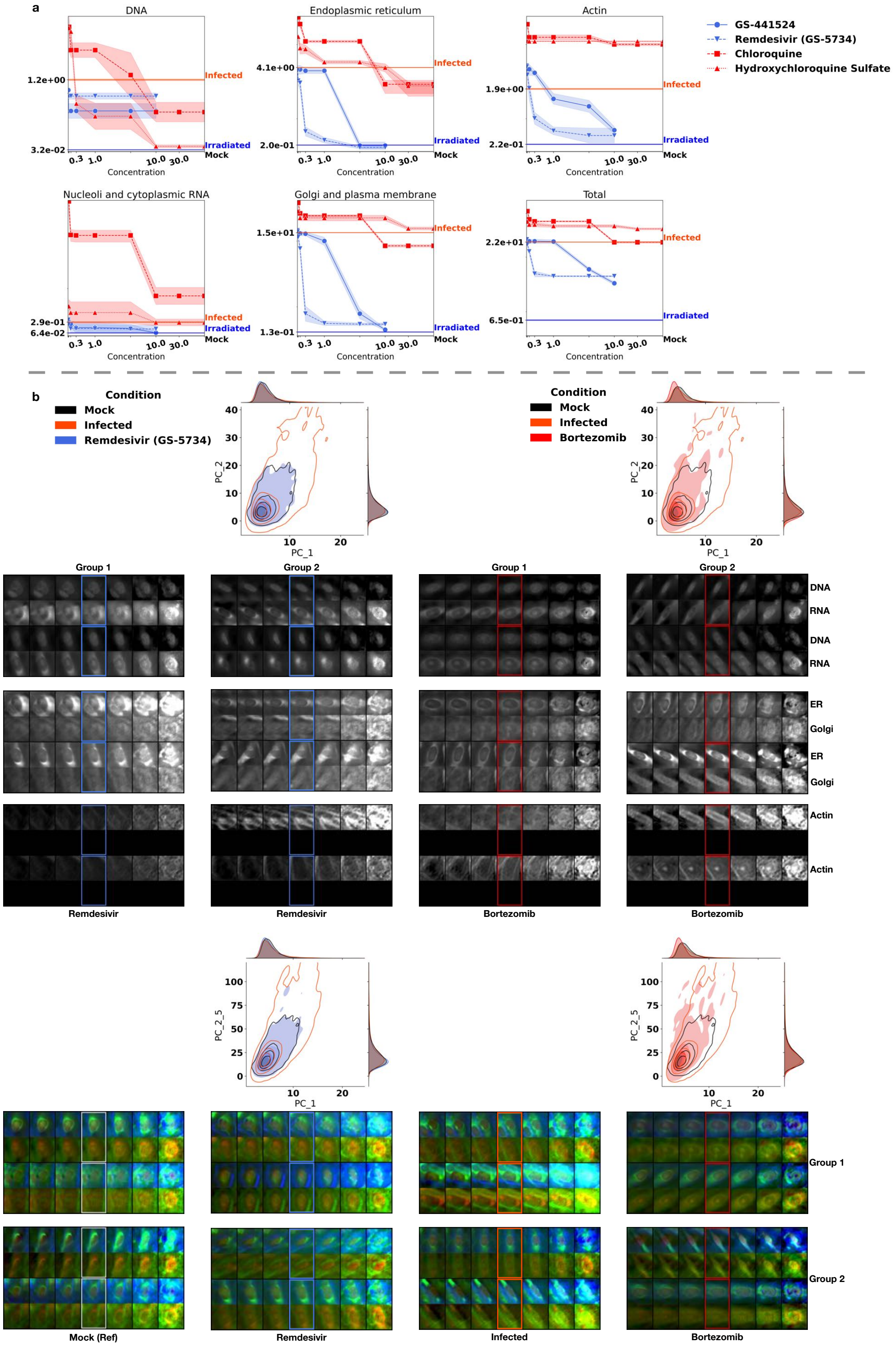


Figure 4: **Identification of drug-concentration dependent effects and visual interpretation for key drugs of interest in the HRCE cell-line**. **a**, The proposed d_{LEA} of different drug concentrations for individual and all fluorescent channels. Here, we report the mean d_{LEA} (with standard deviation) averaged on 4 randomly sampled cell collections. **b**, The PCA plots and phenotypic transitions driven by manipulating the largest (top) and 5 largest (bottom) principal component(s). The bounding box indicates the reconstructed image.

References

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6711–6720, 2021.

Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *arXiv preprint arXiv:2201.13433*, 2022.

Álvaro Avezum, Gustavo BF Oliveira, Haliton Oliveira, Rosa C Lucchetta, Valéria FA Pereira, André L Dabarian, D Ricardo, O Vieira, Daniel V Silva, Adrian PM Kormann, et al. Hydroxychloroquine versus placebo in the treatment of non-hospitalised patients with covid-19 (cope–coalition v): A double-blind, multicentre, randomised, controlled trial. *The Lancet Regional Health-Americas*, 11:100243, 2022.

Amit H Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Or Patashnik, and Daniel Cohen-Or. State-of-the-art in the architecture, methods and applications of stylegan. *arXiv preprint arXiv:2202.14020*, 2022.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

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

Luis Pedro Coelho. Mahotas: Open source software for scriptable computer vision. *arXiv preprint arXiv:1211.4907*, 2012.

Michael F Cuccarese, Berton A Earnshaw, Katie Heiser, Ben Fogelson, Chadwick T Davis, Peter F McLean, Hannah B Gordon, Kathleen-Rose Skelly, Fiona L Weathersby, Vlad Rodic, et al. Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and covid-19 drug discovery. *bioRxiv*, 2020.

Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *arXiv preprint arXiv:2006.05690*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

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. *Advances in neural information processing systems*, 30, 2017.

Biying Hu, Shaoying Huang, and Lianghong Yin. The cytokine storm and covid-19. *Journal of medical virology*, 93(1):250–256, 2021.

Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data, 2009.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Goran Kokic, Hauke S Hillen, Dimitry Tegunov, Christian Dienemann, Florian Seitz, Jana Schmitzova, Lucas Farnung, Aaron Siewert, Claudia Höbartner, and Patrick Cramer. Mechanism of sars-cov-2 polymerase stalling by remdesivir. *Nature communications*, 12(1):279, 2021.

Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pp. 279–292, 1961.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Brandon Malone, Nadya Urakova, Eric J Snijder, and Elizabeth A Campbell. Structures and functions of coronavirus replication–transcription complexes and their relevance for sars-cov-2 drug design. *Nature Reviews Molecular Cell Biology*, 23(1):21–39, 2022.

Donald B Owen. The power of student's t-test. *Journal of the American Statistical Association*, 60 (309):320–333, 1965.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

GL Richard and DL Hahs-Vaughn. Statistical concepts: a second course, 2007.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2287–2296, 2021.

M Roustit, R Guilhaumou, Mathieu Molimard, M-D Drici, S Laporte, J-L Montastruc, French Society of Pharmacology, Therapeutics (SFPT, et al. Chloroquine and hydroxychloroquine in the management of covid-19: much kerfuffle but little evidence. *Therapies*, 75(4):363–370, 2020.

Sultan AM Saghir, Naif A AlGabri, Mahmoud M Alagawany, Youssef A Attia, Salem R Alyileili, Shaaban S Elnesr, Manal E Shafi, Omar YA Al-Shargi, Nader Al-Balagi, Abdullah S Alwajeeh, et al. Chloroquine and hydroxychloroquine for the prevention and treatment of covid-19: A fiction, hope or hype? an updated review. *Therapeutics and clinical risk management*, 17:371, 2021.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1532–1540, 2021.

Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Frank Wilcoxon. Individual comparisons by ranking methods. *Breakthroughs in statistics*, pp. 196–202, 1992.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

Jiqing Wu and Viktor Koelzer. Sorted eigenvalue comparison d_{Eig}: A simple alternative to d_{FID}. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

Jiqing Wu, Nanda Horeweg, Marco de Bruyn, Remi A Nout, Ina M Jürgenliemk-Schulz, Ludy CHW Lutgens, Jan J Jobsen, Elzbieta M van der Steen-Banasik, Hans W Nijman, Vincent THBM Smit, et al. Automated causal inference in application to randomized controlled clinical trials. *Nature Machine Intelligence*, 2022.

Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872, 2021.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

A Appendix

		VERO (Proposed)	VERO (Ensemble)	**HRCE (Proposed)**	HRCE (Ensemble)
DNA	PSNR	35.24 ± 1.87	34.67 ± 2.29	44.32 ± 2.48	38.25 ± 3.66
	SSIM	0.95 ± 0.01	0.95 ± 0.02	0.99 ± 0.01	0.95 ± 0.05
ER	PSNR	35.31 ± 1.94	32.20 ± 2.11	35.27 ± 2.36	29.42 ± 3.12
	SSIM	0.96 ± 0.01	0.93 ± 0.02	0.95 ± 0.02	0.83 ± 0.09
Actin	PSNR	33.54 ± 2.70	31.74 ± 2.72	33.68 ± 3.21	29.22 ± 3.44
	SSIM	0.93 ± 0.02	0.90 ± 0.03	0.92 ± 0.04	0.78 ± 0.09
RNA	PSNR	36.93 ± 1.51	34.59 ± 1.87	39.00 ± 2.56	33.87 ± 3.07
	SSIM	0.97 ± 0.01	0.95 ± 0.02	0.96 ± 0.02	0.90 ± 0.06
Golgi	PSNR	37.23 ± 2.24	33.84 ± 2.08	36.48 ± 2.44	32.69 ± 3.29
	SSIM	0.95 ± 0.02	0.92 ± 0.02	0.95 ± 0.03	0.87 ± 0.07
Total	PSNR	35.27 ± 1.86	33.14 ± 2.11	36.26 ± 2.40	31.36 ± 2.95
	SSIM	0.95 ± 0.01	0.93 ± 0.02	0.95 ± 0.02	0.87 ± 0.07

Table 1: The numerical comparison of reconstruction results between the proposed and ensemble LEA on VERO and HRCE.

A.1 LEA benchmarking

For each fluorescent channel, the proposed LEA outperforms the ensemble variant on both cell-line experiments in terms of overall better peak signal-to-noise Ratio (PSNR) and structural similarity index measure SSIM scores (Appendix Tab. 1). This indicates superior reconstruction quality by the proposed LEA over the ensemble approach (Please see Fig. 2 (a) and (f) for visual results).

Importantly, the proposed LEA is well calibrated by the small distance between mock control and irradiated control cells (Fig. 2 (b,d,g,i)), which is consistent with the expected similar phenotypic and biological characteristics shared by the control conditions 'mock' (cells in culture medium without viral stimulation) and 'irradiated' (cells in culture medium incubated with the inactivated virus). In contrast, such a verifiable calibration cannot be reproduced with the ensemble approach. For the HRCE experiment, Fig. 6 (c, d) in the Appendix shows an inexplicably small difference between mock and infected cell populations as compared to the control conditions, which contradicts the expected phenoprint heterogeneity induced by the virus. In terms of the largest eigenvalues, $p' = 5$ robustifies the drug effect quantification of d_{LEA} and shows an improved consistency with the hit score (Cuccarese et al., 2020), which clearly differs from $p' = 1$, a balanced alternative of d_{Eig} (See Fig. 7 (VERO) and 8 (HRCE) in the Appendix for more detail).

A.2 Toy Results

In reference to Fig. 1(b), here we describe the architecture of LEA in detail and provide insights into the evaluation of model performance. This is done by conducting a separate series of experiments on the HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions (Tschandl et al., 2018). Complementary to above drug screening studies, these experiments support the technical robustness of the LEA pipeline to various types of input data and showcase its domain-agnostic facet. As HAM10000 includes clinical images of human skin lesions, it is easily interpretable by domain experts and allows conceptual validation of the LEA results in the context of well-established disease categories. Concretely, the HAM10000 (Tschandl et al., 2018) dataset has 7 categories of skin lesion images including actinic keratoses (akiec), basal cell carcinoma (bcc), benign keratosis (bkl), dermatofibroma (df), melanocytic nevi (nv), melanoma (mel) and vascular skin lesion (vasc). As the RGB channels of skin lesion images jointly inform the clinical presentation, we train the LEA pipeline on images with all RGB channels simultaneously. For the sake of probing d_{LEA} within such a distinct domain, we design simple interpolation experiments among different categories as follows. Considering the benign nevus ('nv') collection of images as the reference and comparing this to the malignant counterpart (*e.g.*, malignant melanoma, mel), we randomly mix the images of 'nv' and 'mel' (*e.g.*, $\mathbf{x}_{\mathsf{mel},1}, \ldots, \mathbf{x}_{\mathsf{mel},n_{\mathsf{mel}}}, \mathbf{x}_{\mathsf{nv},1}, \ldots, \mathbf{x}_{\mathsf{nv},n_{\mathsf{nv}}}$) according to the interpolation weight $w = \frac{n_{\mathsf{nv}}}{n_{\mathsf{nv}}+n_{\mathsf{mel}}}$. Then, we measure the distribution difference between nv and the mixed collection by d_{LEA}. Ideally, we should observe that d_{LEA} converges to 0 when w is shifted from 0 to 1 with the increasing inclusion of 'nv' ($n_{\mathsf{nv}} \uparrow$) and exclusion of 'mel' images ($n_{\mathsf{mel}} \downarrow$) (Fig. 5 (b)). To avoid the sample imbalance between the reference (nv, 6705 images) and compared categories, we take mel (1113), bkl (1099), and bcc (512) for comparison.

HAM10000		**StyleGAN2_pSp (Default)**	StyleGAN2_e4e	StyleGAN3_pSp	StyleGAN3_e4e
nv	PSNR	30.61 ± 2.20	26.40 ± 1.36	31.60 ± 2.62	30.12 ± 2.44
	SSIM	0.89 ± 0.06	0.85 ± 0.07	0.90 ± 0.06	0.87 ± 0.06
mel	PSNR	28.13 ± 2.39	24.80 ± 1.69	28.60 ± 2.55	27.28 ± 2.53
	SSIM	0.85 ± 0.06	0.80 ± 0.08	0.86 ± 0.06	0.83 ± 0.07
bcc	PSNR	28.95 ± 2.60	25.49 ± 1.58	29.48 ± 2.80	28.28 ± 2.72
	SSIM	0.82 ± 0.07	0.75 ± 0.08	0.83 ± 0.07	0.79 ± 0.07
bkl	PSNR	29.06 ± 2.54	25.50 ± 1.48	29.55 ± 2.65	28.34 ± 2.56
	SSIM	0.83 ± 0.07	0.78 ± 0.08	0.85 ± 0.06	0.81 ± 0.07
df	PSNR	30.01 ± 2.29	25.96 ± 1.27	30.49 ± 2.38	29.34 ± 2.37
	SSIM	0.85 ± 0.06	0.79 ± 0.07	0.86 ± 0.05	0.83 ± 0.06
vasc	PSNR	30.52 ± 2.77	25.78 ± 1.35	31.15 ± 3.14	29.72 ± 2.92
	SSIM	0.88 ± 0.07	0.84 ± 0.08	0.89 ± 0.07	0.86 ± 0.08
akiec	PSNR	27.71 ± 2.09	24.86 ± 1.32	28.21 ± 2.21	27.07 ± 2.12
	SSIM	0.79 ± 0.06	0.71 ± 0.08	0.80 ± 0.06	0.76 ± 0.07
Total	PSNR	29.98 ± 2.49	26.01 ± 1.54	30.81 ± 2.88	29.40 ± 2.71
	SSIM	0.87 ± 0.07	0.83 ± 0.08	0.88 ± 0.06	0.85 ± 0.07

Table 2: The numerical comparison of reconstruction results among different model architectures on HAM10000.

A.2.1 Model architecture

Motivated by the impressive achievements of GAN inversion (Bermano et al., 2022), we instantiate LEA with the state-of-the-art GAN inversion model (Fig. 1(b)). Firstly, we learn the decoder (generator) under the StyleGAN (Karras et al., 2019; 2021) framework, which has proved to be successful in hallucinating high-quality natural images. Based on the training protocols suggested in the widely-used repositories[2][3], we pre-train the StyleGAN2 (Karras et al., 2019) and StyleGAN3 (Karras et al., 2021) on HAM10000 to obtain such a decoder (generator) that can synthesize faithful skin lesion images (Please see the Appendix Fig. 10). Next, we launch the encoder training to learn robust latent representations for image reconstruction. Specifically, we apply two practical architectures 'encoder for editing' (Tov et al., 2021) (e4e) and 'pixel2style2pixel' (Richardson et al., 2021) (pSp) for comparison, both of which start with a ResNet backbone and then concatenate a feature pyramid network (Lin et al., 2017). Similar to the loss design of these studies that enable high-fidelity image reconstruction, we determine our objective to be $\mathcal{L} = \gamma_1 \mathcal{L}_{\text{moco}} + \gamma_2 \mathcal{L}_2$, where $\mathcal{L}_{\text{moco}}$ is the contrastive loss that is superior in visual representation learning (He et al., 2020), and $\mathcal{L}_2$ is the l_2 reconstruction loss. Eventually, we report the quantitative reconstruction results in Tab. 2 among compared architectures.

A.2.2 pSp VS e4e

In terms of quantitative scores such as PSNR and SSIM, it is clear to see that the pSp encoder in combination with either StyleGAN2 or 3 decoder outperforms e4e by a clear margin, suggesting superior image reconstruction qualities. This can also be verified by the image samples presented in Fig. 5 (a), the images reconstructed from the representations of pSp encoder reserve finer detail of lesion demarcation and skin pigmentation, while the e4e encoder tend to produce more blurry reconstructions. As a result, we take the pSp architecture as the default encoder.

A.2.3 StyleGAN2 VS StyleGAN3

Following the encoder architecture comparison, we investigate the variants of StyleGAN decoder. While StyleGAN3_pSp achieves better PSNR and SSIM scores than StyleGAN2_pSp, the qualitative difference in image reconstruction appears marginal (Fig. 5 (a)). Furthermore, when examining the d_{LEA} behavior with the increasing interpolation weight, we found notable differences between the two architectures. Compared to StyleGAN2, Fig. 5 (b) shows that d_{LEA} computed with StyleGAN3_e4e increases unexpectedly from $w = 0.25$ to $w = 0.5$ for the mixture data collection of nv and mel images, which is in conflict with the fact that more inclusion of benign mole images should reduce the distance to the nv reference category. With regards to nv, d_{LEA} of StyleGAN3_pSp surprisingly suggests a larger data difference of bkl (bcc) than mel. This is also counter-intuitive as malignant neoplasm is more likely to present alert appearances in lesion size and pigmentation heterogeneity that clearly differentiate itself from the benign mole. Besides, we notice that the learned representations of StyleGAN3 tend to be more convoluted and are thus more challenging to support clear biological interpretation (See for example Appendix Fig. 11). Since StyleGAN3 is motivated by the texture-sticking drawback occurring in natural images and imposes equivariant translation and rotation on learned representations (Karras et al., 2021), it may explain why StyleGAN3 does not adapt well to biomedical images from substantially distinct modalities. This is also reflected by the drawbacks identified by Alaluf et al. (2022) for natural images.

In summary, we consider **StyleGAN2_pSp** the default architecture for conducting drug screening and skin lesion experiments through this article.

A.2.4 Further comparisons

Next, we investigate the d_{LEA} performance with regards to the amount p' of the largest eigenvalues utilized in Eq. 2. As we can see in Fig. 5 (b), d_{LEA} shows comparable decreasing scores with the increasing weight of including more nv images for $p' = 1, \ldots, 4, 5$. Such results demonstrate the feasibility and robustness of d_{LEA} computed with the largest eigenvalues for the RGB imaging. In combination with the multiplexed imaging for drug screening, we narrow down the amount p' of

[2] StyleGAN2: `https://github.com/rosinality/stylegan2-pytorch.git`

[3] StyleGAN3: `https://github.com/NVlabs/stylegan3.git`

the largest eigenvalues to 5 in this paper. In addition, we compare d_{LEA} with the sota statistical tests and widely used scalar-valued scores. For the former, we report the (average) p-values computed with two collections of eigenvalues. In the meantime, we compute d_{KID} with multiple subsets of randomly sampled latent representations and d_{Eig} with two SCMs (Def. 1). Regarding statistical tests, we have witnessed either contradictory behaviors obtained by F_test or uninformative Levene_test and Wilcoxon_test, which confirms the challenging adaption to high-dimensional cases, as discussed in the introduction. Although meaningful decreasing curves can be obtained with d_{KID}, it shows clear fluctuations with large standard deviations. This is mainly due to the additional randomness that comes from subset selection, which is not present in other measurements. Without the imbalance issue introduced in the skin lesion dataset, d_{Eig} shows plausible decreasing trends similar to d_{LEA}.

A.2.5 Clinical Interpretations

As shown in Fig. 5 (c), clear patterns in terms of lesion size emerge in two different groups clustered with k-means. This has been persistently presented among different categories. If we manipulate the largest principal component (top rows of Fig. 5 (c)) with increasing $\beta^1_{i,k}$ (from left to right), the 'nv' images show a poor lesion demarcation, further increase in lesion size and pigmentation heterogeneity, with the lesions displayed the right-most images highlights clear pathological changes towards a clinically suggestive appearance of malignancy. Considering the largest eigenvalue ($\times 10^5$) 4.41 for nv versus 6.04 for mel, the appearance shift towards malignancy by enlarging the principal component of nv representations can indeed explain the eigenvalue difference $4.41 < 6.04$. Comparable observations can be also made when investigating the 5 largest principal components. Apart from similar lesion size patterns arising from the k-means clustering, we show some distinct samples clustered in the two groups (Bottom rows of Fig. 5 (c)). Accordingly, the PCA plots regarding the 5 largest eigenvalues verify the distinguished yet consistent heterogeneity quantification among mel, bcc, bkl and nv.

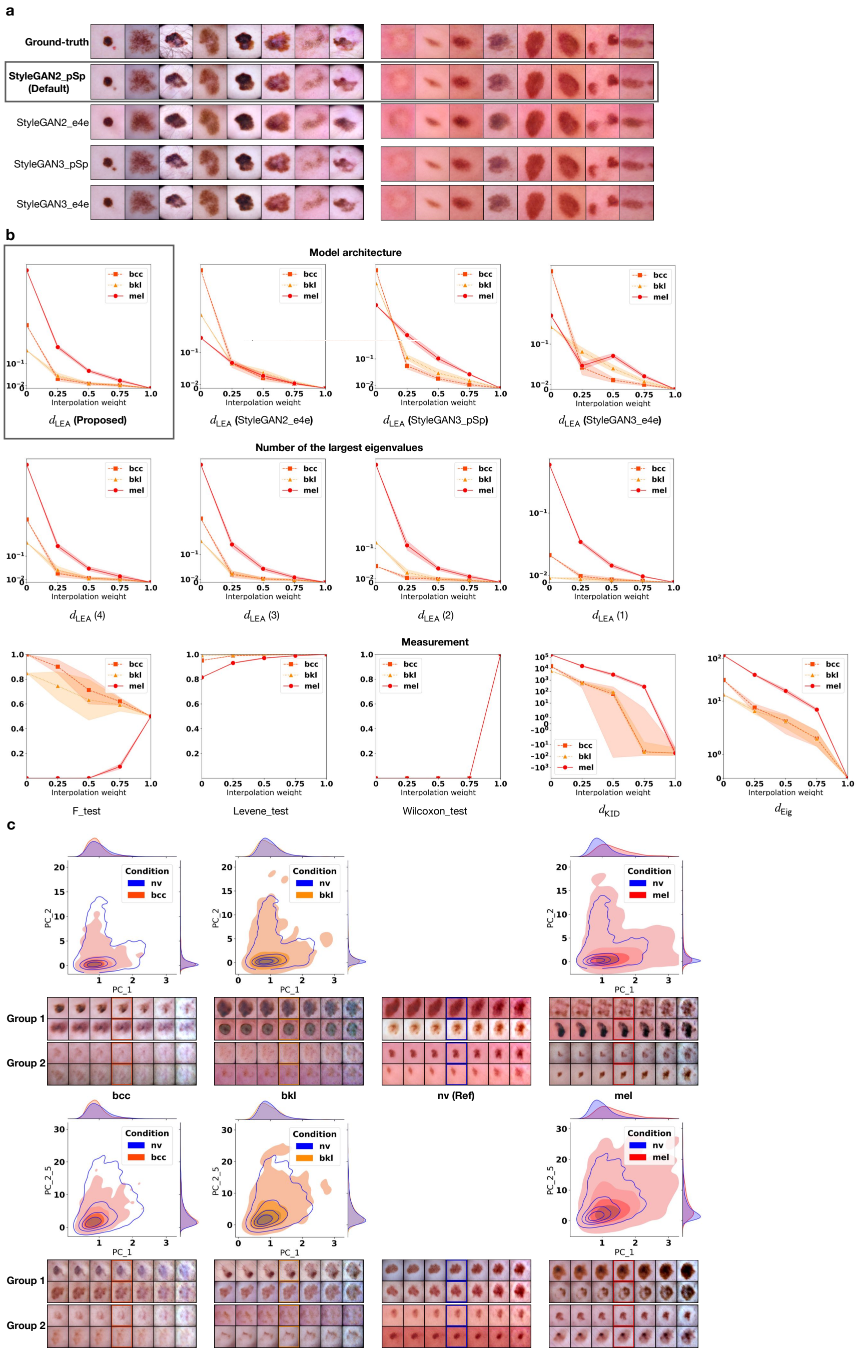


Figure 5: **The data interpolation quantification and visual interpretation on HAM10000 experiments**. **a**, The reconstructed samples obtained by 4 different architectures. **b**, The d_{LEA} comparison of data interpolation regarding different architectures, number of the largest eigenvalues, and existing measurements. Here, we report the mean d_{LEA} and compared measurements (with standard deviation) averaged on 4 randomly sampled data mixtures given the interpolation weight. **c**, The PCA plots and morphological transitions driven by manipulating the largest principal components. The bounding box indicates the reconstructed image.

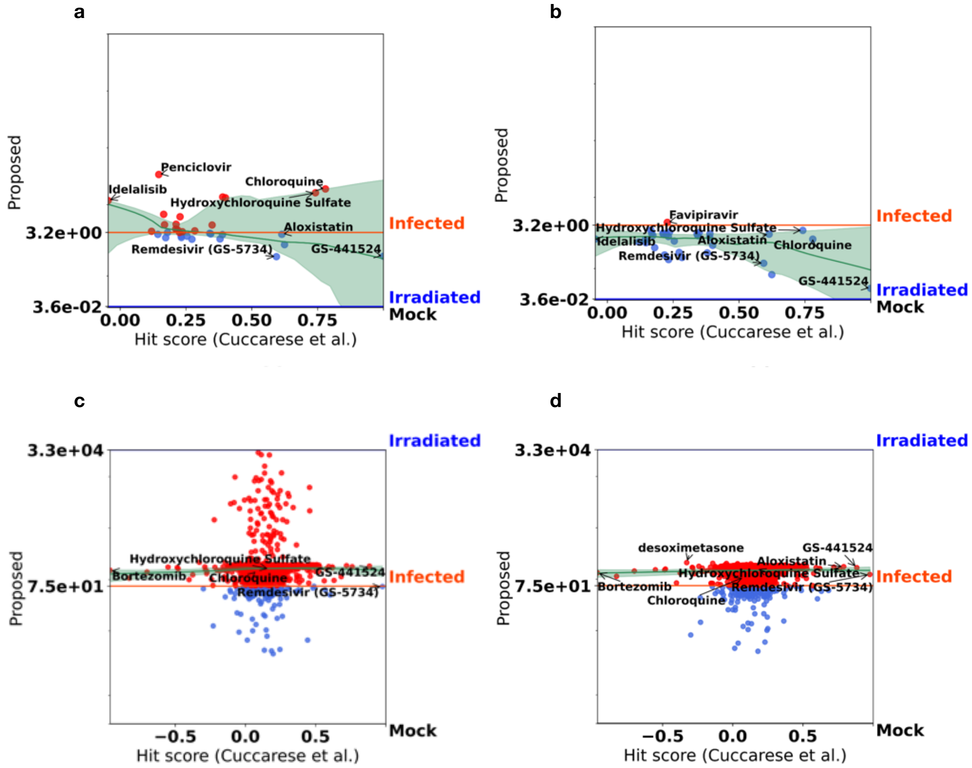
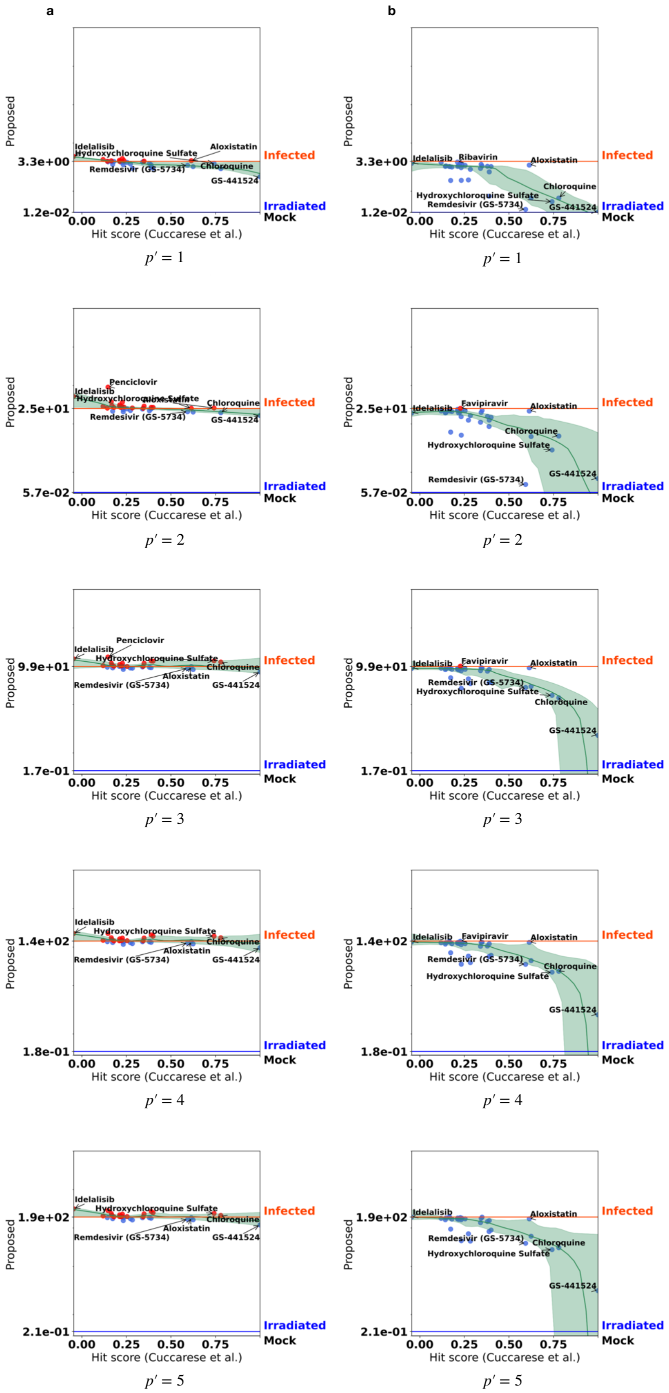


Figure 6: **Quantification comparison of cell-based COVID-19 drug responses between the baseline (Cuccarese Cuccarese et al. (2020)) and d_{LEA} (Ensemble).** **a** (VERO) and **c** (HRCE): The quantitative comparison between the hit score (Cuccarese et al., 2020) and d_{LEA} (Ensemble) with the latent representations of all drug concentrations. **b** (VERO) and **d** (HRCE): The quantitative comparison between the hit score and d_{LEA} with the latent representations of optimal drug concentration.



$p' = 1$

$p' = 1$

$p' = 2$

$p' = 2$

$p' = 3$

$p' = 3$

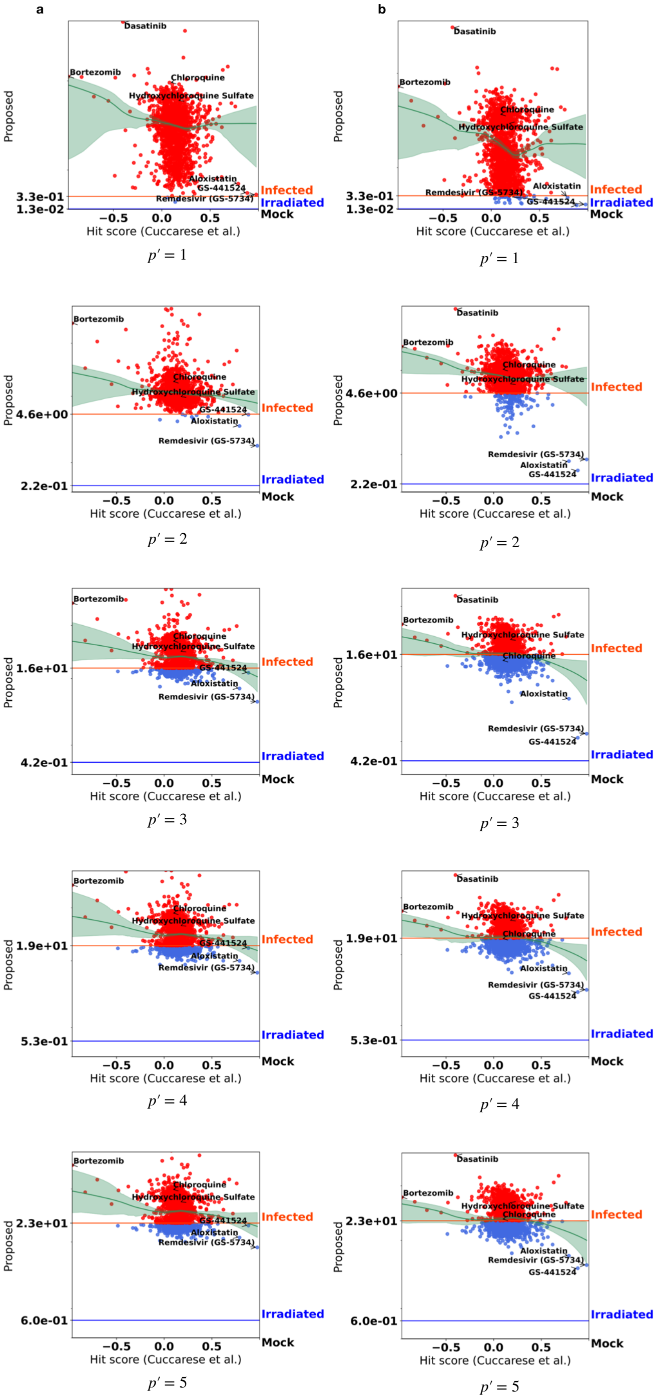
$p' = 4$

$p' = 4$

$p' = 5$

$p' = 5$

Figure 7: **Quantification result d_{LEA} of VERO *w.r.t.* different amount of the largest eigenvalues**. **a**: d_{LEA} computed with the latent representations of all drug concentrations. **b**: d_{LEA} computed with the latent representations of optimal drug concentration.



$p' = 1$ $p' = 1$

$p' = 2$ $p' = 2$

$p' = 3$ $p' = 3$

$p' = 4$ $p' = 4$

$p' = 5$ $p' = 5$

Figure 8: **Quantification result d_{LEA} of HRCE *w.r.t.* different amount of the largest eigenvalues**. **a**: d_{LEA} computed with the latent representations of all drug concentrations. **b**: d_{LEA} computed with the latent representations of optimal drug concentration.

		HUVEC (Proposed)	HUVEC (Ensemble)
DNA	PSNR	44.70 ± 3.24	41.03 ± 3.25
	SSIM	0.99 ± 0.02	0.98 ± 0.02
ER	PSNR	38.11 ± 3.27	36.90 ± 3.80
	SSIM	0.97 ± 0.02	0.95 ± 0.03
Actin	PSNR	45.04 ± 2.68	42.68 ± 2.72
	SSIM	0.98 ± 0.01	0.97 ± 0.02
RNA	PSNR	42.28 ± 2.60	41.93 ± 2.97
	SSIM	0.98 ± 0.01	0.97 ± 0.02
Mitochondria	PSNR	45.11 ± 2.20	42.93 ± 2.04
	SSIM	0.98 ± 0.01	0.98 ± 0.02
Golgi	PSNR	42.02 ± 2.78	39.59 ± 2.95
	SSIM	0.98 ± 0.01	0.96 ± 0.03
Total	PSNR	41.83 ± 2.66	40.03 ± 2.95
	SSIM	0.98 ± 0.01	0.97 ± 0.02

Table 3: The numerical comparison of reconstruction results among different model architectures on HUVEC.

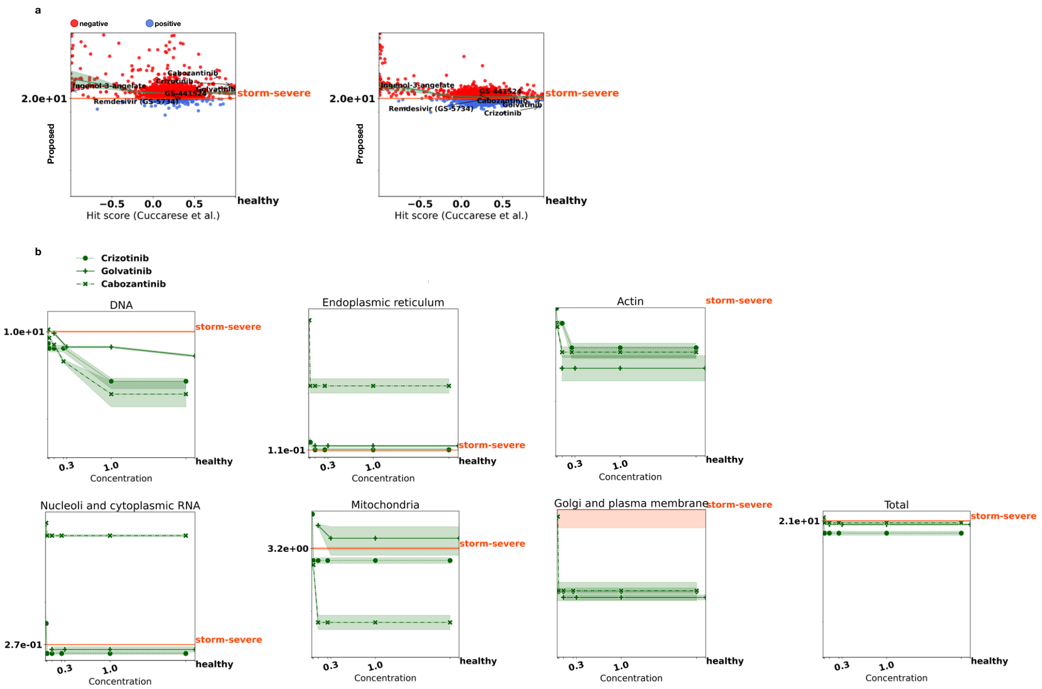


Figure 9: **The quantification results of HUVEC experiment**. **a**, The quantitative comparison between the hit score (Cuccarese et al., 2020) and d_{LEA} with latent representations of all drug concentrations (left) and the optimal drug concentration (right). Our drug effects (positive/negative) are thresholded by the d_{LEA} of storm-severe cells without drug treatment. **b**, The proposed d_{LEA} of different drug concentrations for individual and all fluorescent channels. Here, we report the mean d_{LEA} (with standard deviation) averaged on 4 randomly sampled cell collections.

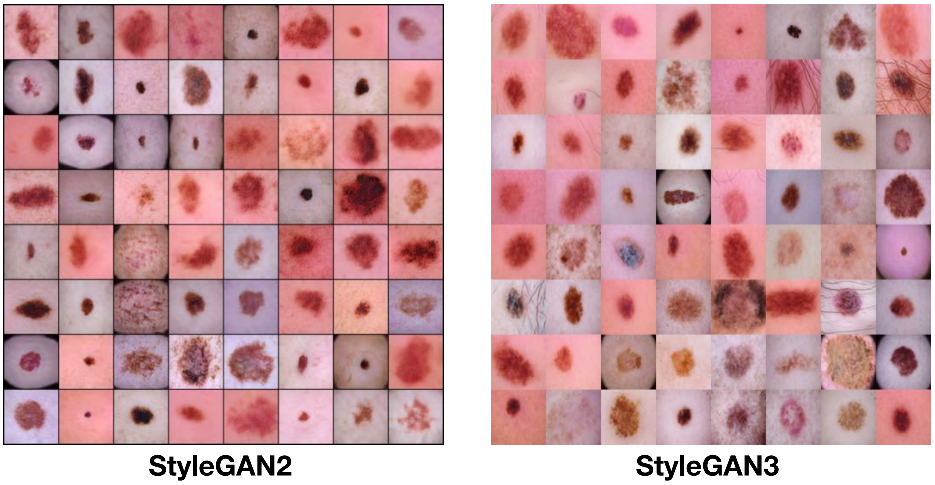


Figure 10: **The non-existent skin lesion images synthesized by StyleGAN2 (left) and StyleGAN3 (right)**.

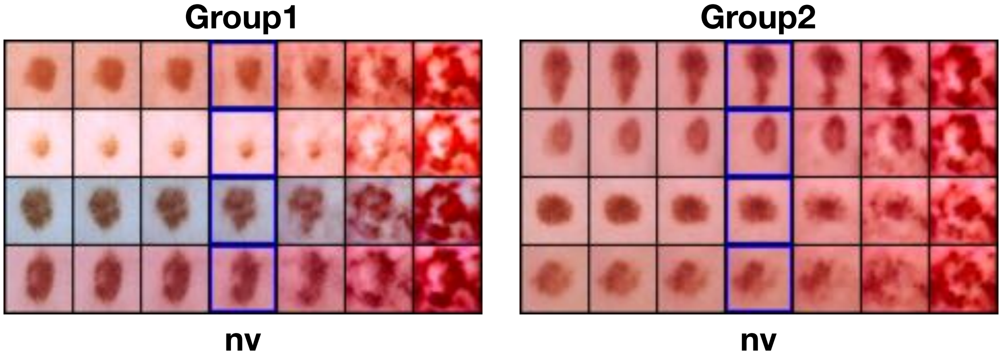


Figure 11: **The phenotypic transitions driven by manipulating the largest principal component, which are derived from the latent representations of StyleGAN3_psp**.