Multi-ContrastiveVAE disentangles perturbation effects in single cell images from optical pooled screens

Zitong Jerry Wang[1,†], **Romain Lopez**[2,3], **Jan-Christian Hütter**[2], **Takamasa Kudo**[2], **Heming Yao**[2], **Philipp Hanslovsky**[2], **Burkhard Höckendorf**[2], **Aviv Regev**[2]

[1]California Institute of Technology
[2]Genentech Research and Early Development
[3]Stanford University
[†]Work performed during an internship at Genentech, Inc.

Abstract

Optical pooled screens (OPS) enable unbiased and cost-effective interrogation of gene function by generating images of millions of cells across thousands of perturbations. However, the analysis of OPS data remains a hurdle because it still mainly relies on hand-crafted features, which can be difficult to deploy across complex data sets. Additionally, most unsupervised feature extraction methods based on neural networks (such as auto-encoders) have difficulty isolating the effect of perturbations from the natural variations among cells. We therefore propose a contrastive analysis framework that is more effective at disentangling the phenotypes induced by perturbation from natural cell-cell heterogeneity present in an unperturbed cell population. By analyzing a significant data set of over 30 million cells across more than $5,000$ genetic perturbations, we demonstrate that our method significantly outperforms traditional methods in generating biologically-informative embeddings and mitigating technical artifacts. Furthermore, the interpretable part of our model enables us to pinpoint perturbations that generate novel phenotypes from the ones that only shift the distribution of existing phenotypes. Our approach can be readily applied to other small-molecule and genetic perturbation data sets with highly multiplexed images, enhancing the efficiency and precision in identifying and interpreting perturbation-specific phenotypic patterns, paving the way for deeper insights and discoveries in OPS analysis.

1 Introduction

Large-scale, pooled genetic perturbation screens in cells enable unbiased interrogation of gene function [1]. Optical pooled screens (OPS) employ cell imaging as phenotypic read-out for characterizing perturbation effects as it is high-throughput and low-cost [2].

Traditional phenotype analysis requires extracting from each image a set of hand-crafted morphological features [3, 4]. Conventional tools like CellProfiler (CP) usually apply a set of pre-defined filters to images of individual cells in order to summarize the data from each single cell into a data point using thousands of features. There are several limitations to these approaches. First, because the filters have been engineered on previous data sets, with different cell types and experimental conditions, hand-crafted features may be inflexible to capture novel morphological phenotypes. Indeed, because cellular morphology drastically changes across contexts and biological systems, the set of optimal features may be different for each experiment. Second, those extracted feature are limited in their ability to capture interactions between channels. For example, CP accomplishes this by applying filters to pairs of channels (usually for images with two to five channels). However, this will be difficult to scale in order to capture more complex patterns in images with tens to hundreds of channels which are becoming increasingly common [4].

Preprint. Under review.

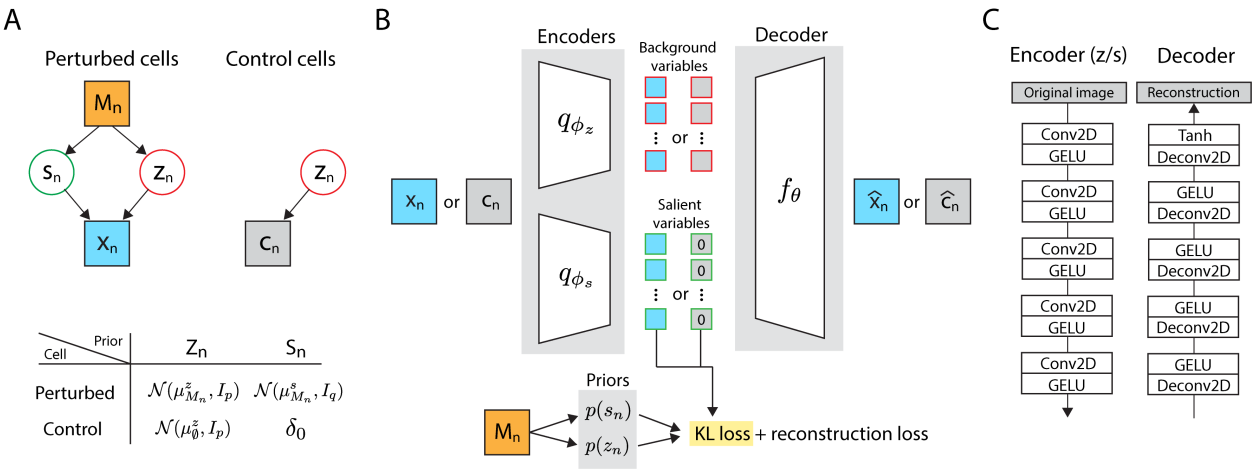


Figure 1: **A contrastive analysis framework for analyzing single cell images from optical pooled screens.** (A) Generative model for perturbed and control cells. The table shows the prior distributions used for the background and salient variables of the perturbed and control population. (B) Schematic of the Multi-Contrastive variational autoencoder (mcVAE) framework. (C) Neural network architectures of the encoder and the decoder.

Advances in deep learning could overcome limitations of hand-crafted features by learning representations directly from data [3, 5]. In particular, the Variational Auto-Encoder (VAE) is a powerful deep generative framework to capture latent structure in complex data distributions in an unsupervised manner [6, 7]. However, in the context of learning perturbation effects from images of cells, standard VAE implementations suffer from the important drawback that it can be difficult to isolate perturbation effects from natural cell-cell heterogeneity (e.g., due to stages of the cell cycle), which can exhibit much greater variation compared to the phenotypic effect of perturbations.

The contrastive analysis (CA) framework offers a potential solution to identify and isolate patterns induced by perturbations using a background data set to remove those natural variations [8]. In our setting, the background data set is composed of images of control cells that have not been perturbed. For example, contrastive principal components analysis (cPCA) seeks to identify salient principal components in a target data set by identifying linear combinations of features that are enriched in that data set relative to a background data set, rather than those that simply have the most variance [9]. Recently, CA methods based on neural networks have demonstrated effectiveness in discovering nonlinear latent features that are enriched in one dataset compared to another [10–14]. These approaches often assume that only two datasets are being processed, both stemming from identically and independently distributed data distributions: the background distribution and the target distribution. This approach is limited in that it does not explicitly model each perturbation as a unique distribution for comparison with the background distribution.

We therefore propose Multi-ContrastiveVAE, a CA framework for the setting of comparing multiple data sets to a reference data set, with a specific architecture tailored for cell imaging data sets from optical pooled screens (OPS). We applied our method to a large-scale imaging dataset comprising over 30 million cells across more than $5,000$ genetic perturbations, as detailed in [15]. Our approach more accurately identifies perturbation-specific phenotypes compared to non-contrastive methods, effectively distinguishing them from cell-to-cell variations that persist across perturbations. Specifically, our method effectively separates multiple sources of technical artifacts from single-cell images including non-biological variations due to batch effects, uneven plating, and uneven FOV illumination. Furthermore, our method disentangles perturbation effects into separate latent spaces depending on whether the perturbation induces novel phenotypes unseen in the control cell population. Lastly, by comparing the embeddings from the two latent spaces, we can more effectively identify genes with established roles in mitosis. Our approach is readily applicable to other perturbation data sets including both drug and genetic perturbation, with highly multiplexed images.

2 Methods

We developed the Multi-Contrastive Variational Autoencoder (mcVAE) to disentangle perturbation effects from natural cell-to-cell variations in large-scale perturbation data sets. This model extends the CA framework, allowing for the comparison of multiple groups against a single reference group (the background data set composed of control, unperturbed cells). Our framework is based on the generative model illustrated in Figure 1A, and builds on recent work on CA [10, 14].

Generative Model For each cell n, we observe the perturbation label $M_n \in \{\emptyset, 1, \ldots, K\}$, where $\emptyset$ denotes a non-targeting control (NTC) perturbation (i.e., no effect) and K denotes the number of distinct perturbations. Let n be for now a perturbed cell, that is $M_n \neq \emptyset$. Let latent variable

$$z_n \sim \text{Normal}\left(\mu^z_{M_n}, I_p\right)$$

be a low-dimensional random vector encoding cellular variations that naturally exist in the control cells (the background data set). The mean of the prior $p(z_n \mid M_n)$ varies with the perturbation label M_n to account for the fact that perturbations may shift the density of cells towards certain preexisting cellular states from the control population. We refer to z as the *background* latent space, or embedding. Then, let latent variable

$$s_n \sim \text{Normal}\left(\mu^s_{M_n}, I_q\right)$$

be a low-dimensional vector encoding variations due to perturbations. The mean of the prior $p(s_n \mid M_n)$ is shifted by $\mu^s_{M_n}$ to account for the fact that different perturbations may incur different changes in the data distribution. We refer to s as the *salient* latent space, or embedding. All the images $x_n \in \mathbb{R}^d$ have the same number of pixels d. We assume that each pixel j in each image x_{nj} is generated as:

$$x_n \sim \text{Normal}\left(f_\theta(z_n, s_n), \sigma^2 I_d\right),$$

where f_θ is a neural network taking value in the hypercube $[-1, 1]^d$.

In order to break symmetry between latent variables s and z, we exploit the control cells. The data distribution for the images of the control cells (i.e., $M_n = \emptyset$) is the result of an intervention

$$p\left(x_n \mid M_n = \emptyset\right) = p\left(x_n \mid do(s_n = 0), do(\mu^z_{M_n} = \mu^z_\emptyset)\right),$$

where $\mu^z_\emptyset$ denotes the mean of the prior embedding for the embedding of the control cells (it could be set to zero without loss of generality). This assumption is classical in CA, and helps enforce the semantic that only z (the remaining latent variable) may be used to describe data from the control cells.

Interpretation and Significance Departing from established CA models, mcVAE includes additional parameters μ^s and μ^z that capture the heterogeneity of the perturbations. The former captures the fact that perturbations could induce novel phenotypes (to be captured by the salient variables). The latter biases cell states after perturbation towards phenotypes that already existed in the natural population. This represents a significant conceptual departure from the original framework where the background space was interpreted as containing only uninteresting variation while the perturbation effect resides solely within the salient space.

Variational Inference The marginal probability of the data $p(x \mid M)$ is intractable. We therefore proceed to posterior approximation with variational inference to learn the model's parameters. In particular, we use a mean-field variational distribution:

$$\bar{q} = \prod_{M_n = \emptyset} q_{\phi_z}(z_n \mid x_n)\delta_0(s_n) \prod_{M_{n'} \neq \emptyset} q_{\phi_z}(z_{n'} \mid x_{n'}) q_{\phi_s}(s_{n'} \mid x_{n'}).$$

As in the VAE framework [6], each $q_{\phi_z}(z \mid x)$ and $q_{\phi_s}(s \mid x)$ follows a Gaussian distribution with a diagonal covariance matrix. We optimize a composite objective function, corresponding to the sum of the evidence lower bound (ELBO) for perturbed cells, and for control cells.

Adapting the work of [10], the ELBO for a perturbed cell n is derived as:

$$\log p(x_n \mid M_n) \geq \mathbb{E}_{q_{\phi_z}(z_n|x_n) q_{\phi_s}(s_n|x_n)} \log p_\theta\left(x_n \mid z_n, s_n\right) \tag{1}$$

$$- \text{KL}\left(q_{\phi_s}\left(s_n \mid x_n\right) \| p\left(s_n \mid M_n\right)\right) \tag{2}$$

$$- \text{KL}\left(q_{\phi_z}\left(z_n \mid x_n\right) \| p\left(z_n \mid M_n\right)\right). \tag{3}$$

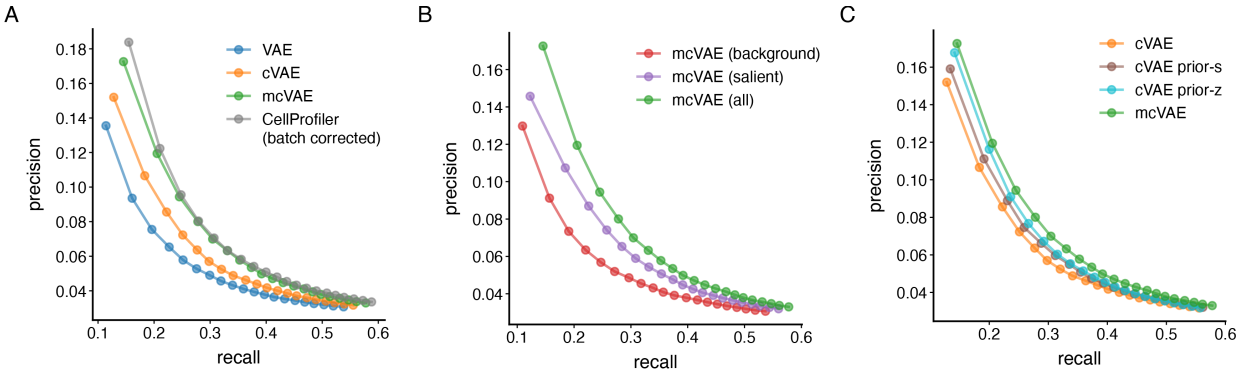


Figure 2: **Multi-Contrastive VAE outperforms contrastiveVAE in protein complex identification.** (A) Precision-recall curve of CORUM identification task for vanilla VAE, contrastiveVAE (cVAE), multi-contrastiveVAE (mcVAE) and using CP features; all four embedding spaces have the same dimension of 64. (B) Precision-recall curve when using background alone, salient alone, or a concatenation of both sets of latent variable. (C) Precision-recall curve for cVAE, cVAE with perturbation label fed into just the salient prior, cVAE with perturbation label fed into just the background prior, and our mcVAE where perturbation labels are fed into both the salient and background prior

Similarly, the ELBO for a unperturbed cell n is derived as:

$$\log p(x_n \mid M_n = \emptyset) \geq \mathbb{E}_{q_{\phi_z}(z_n|x_n)} \log p_\theta \left(x_n \mid z_n, 0\right) - \mathrm{KL}\left(q_{\phi_z}\left(z_n \mid x_n\right) \| p\left(z_n \mid \emptyset\right)\right). \tag{4}$$

We summarized those computations in the schematic from Figure 1B.

Architecture and Training Details For the neural networks used in the generative model, and the amortization of the approximate posterior, we use a simple convolutional architecture as shown in Figure 1C (five convolutional layers for the approximate posterior, and five transpose-convolutional layers for the generative model). We used the Adam optimizer [16] with a learning rate of 10^{-4}. Each mini-batch was constructed to contain an equal number of perturbed and control cells. We applied the inverse hyperbolic sine transformation as well as normalization to the training images during pre-processing. All models presented in this paper were trained for 2 epochs with early stopping based on the validation loss. The dimensions of z and s were set to 32 for all results presented in this paper. We regularized the model using the Wasserstein regularization described in [14] to encourage independence between salient and background latent features.

3 Results

We apply our framework to a public imaging data set from a large-scale CRISPR-knockout screening experiment that profiled 31 million HeLa cells affected by $5,000$ distinct genetic perturbations [15]. Four single guide RNA (sgRNA) sequences were selected for each gene target, together with 250 "non-targeting" sgRNA that do not target any gene. For 31 million cells, there is a median of around $6,000$ cells per gene target across each set of four guides. To obtain single-cell images, we extracted 100×100 pixel crops and perturbation label assignments for each cell as described in Appendix A.1. We assess how well the embedding from our mcVAE reflect known biology, how well it isolates for technical artifacts, and how the salient and background space differ in terms of the perturbation effects they capture.

3.1 Assessment of Embeddings Quality from Database of Protein Complexes

We evaluated the effectiveness of our learned embeddings in capturing known biological structures by employing them to predict established protein complexes [17]. Given that perturbations to different subunits of a protein complex are likely to generate analogous cellular phenotypes, it is reasonable to expect the images corresponding to perturbations of genes belonging to the same complex to be closer in embedding space compared when the perturbed genes do not belong to the same complex.

To conduct this assessment, we utilized the CORUM database [18], the most extensive publicly available collection of manually curated mammalian protein complexes as ground truth. By predicting all aggregated gene embeddings up to a given cut-off as true relationships, we aimed to predict which gene pairs co-occur in a CORUM cluster. We subsequently compare the precision-recall curve of various methods, generated by setting different distance thresholds for determining whether two genes are part of the same complex, to evaluate the classification performance.

We compare the performance of mcVAE to three baseline models: standard VAE, contrastive VAE (cVAE), and CellProfiler (CP). Both the standard VAE and cVAE consists of the same encoder and decoder architecture as shown in Figure 1C. The standard VAE has a single encoder and decoder, and the cVAE consists of two encoder and a shared decoder but does not use the perturbation label to adjust the priors of its latent variables. To keep the dimension of the latent space consistent across all models, we do not use all CP features, and instead we first perform PCA at the cell-level and then take the first 64 PCs (77% variance explained).

The standard VAE trailed in performance, being markedly surpassed by the contrastive VAE (cVAE) (Figure 2A). Taking it a step further, our mcVAE not only exceeded the performance of the cVAE but did so to a degree comparable to the improvement seen previously going from the standard VAE to the cVAE, thereby matching the performance of CP.

We further evaluated how each of the background and salient embedding of mcVAE performs in identifying protein complexes (Figure 2B). Interestingly, the best performance for CORUM identifiability was not achieved by discarding the background information; instead, it was attained by concatenating the salient features with the background. Thus it is not surprising that we obtain sub-optimal performance when we remove either the label information from the salient prior or the background prior (Figure 2C). These findings emphasize the critical role that both the salient and background latent variables play in identifying protein complex interactions, and highlights the strength of mcVAE in elucidating complex biological relationships.

3.2 Performance at Disentangling Multiple Sources of Technical Artifacts

Multi-Contrastive VAE automatically isolates multiple, intricate technical artifacts found in cell images without any prior information. First, batch-to-batch variations can emerge from multiple factors, such as changes in culture conditions and staining conditions (Figure 3A, top). As a result, the background embeddings of cells cluster by batch in the UMAP projection, but cells from different batches are well-mixed in the salient space, indicating the salient space is nearly free of batch effect. Next, uneven illumination of a field of view can cause cells near the center to appear brighter than those near the edge. Background embedding of cells capture this variation, while the salient space appears well-mixed, indicating removal of this technical variation (Figure 3A, middle). Lastly, background cell embeddings are separated based on their position in a well, influenced by uneven cell density affecting cell shape and size, while this source of variation is again absent in the salient space (Figure 3A, bottom). Note that for both the position in FOV and well, the corresponding UMAPs are only showing cells from a single batch to better illustrate these additional variations beyond batch effect.

To quantify the presence of technical variation in different embedding spaces, logistic regression models were trained to predict various technical covariates from the cell embeddings (Figure 3B). The salient embeddings are mostly free of technical artifacts, as evident from having the poorest prediction performance, measured by the F1 score and area under receiver operating characteristic curve (AUROC). In contrast, both the background and CP embeddings contain significant technical variations in terms of FOV position and well position. The CP embedding used was batch-corrected by standardizing against the NTC in the corresponding batch. While effective at removing batch effects, such correction was unable to address the other more intricate sources of technical artifacts.

3.3 Background and Salient Embeddings Excel at Predicting Distinct Gene Functions

Though both salient and background embeddings can accurately classify gene functions, their performance varies significantly based on the functional group. We computed a UMAP projection of guide embeddings (cells aggregated to the level of CRISPR guides) in the combined salient-background space, colored by the 17 functional groups which the perturbed gene belongs to assigned in [15] (Figure 4A). Note that genes of the same functional groups tend to cluster together, suggesting that the embedding space is rich in biological information.

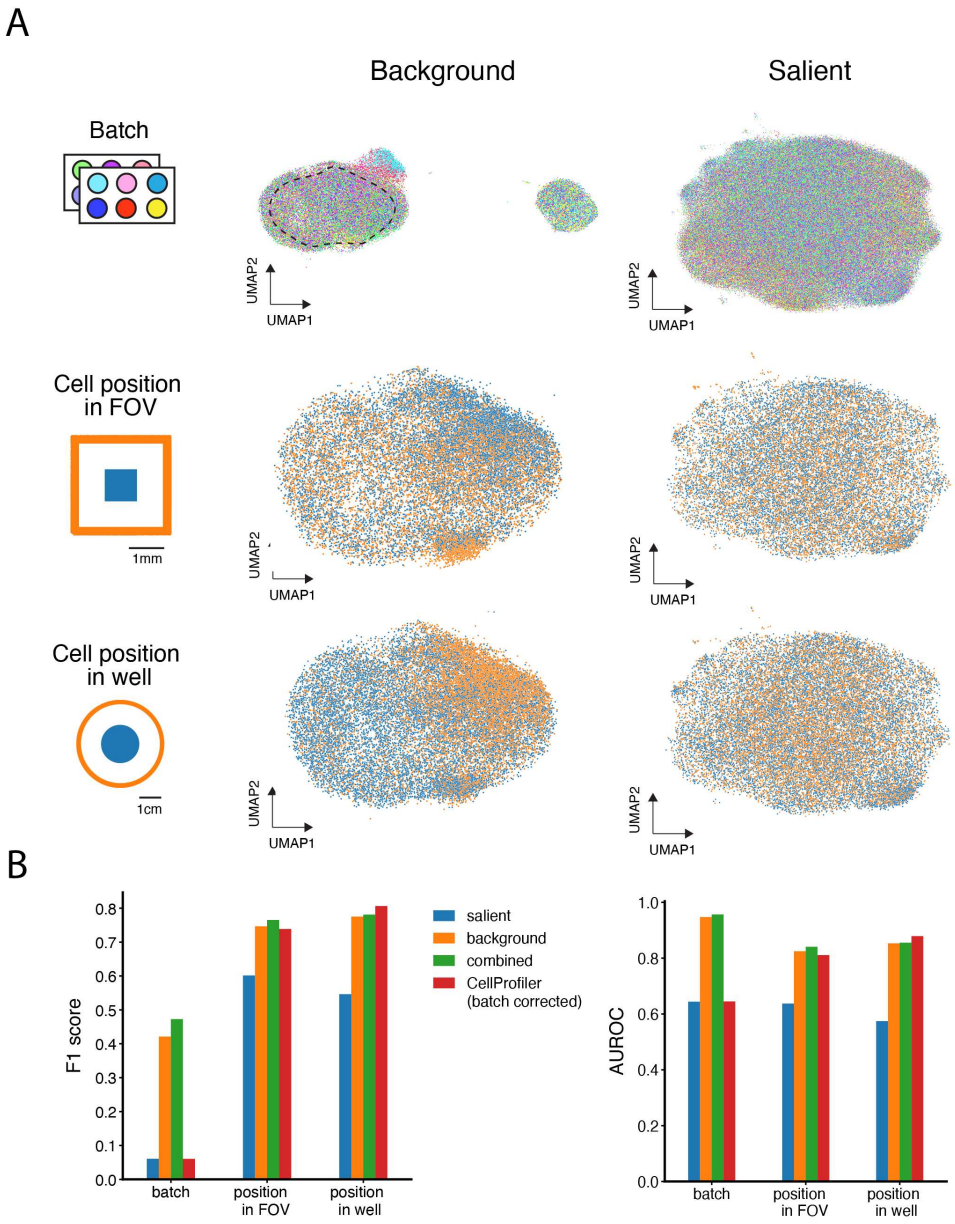


Figure 3: **Multiple experimental technical artifacts are found in both the background space, and CellProfiler features, while the salient space is nearly free of these artifacts.** (A) UMAP projections of background and salient embedding for individual cells colored by their batch (well), position in an image's field of view (edge/center), and position in a well (edge/cell). The embedding for the FOV and well position are for cells from one particular batch, which is also demarcated by dashed black line in the top left UMAP. (B) Performance metrics for a logistic regression model trained to predict the technical covariates from panel A using the cell embeddings from salient space, background space, salient-background concatenated, and batch-corrected CP features.

Salient and background spaces each excel at predicting different gene functions. The confusion matrices (Figure 4B) obtained by training logistic regression models to classify the perturbed gene into one of the functional groups from different guide-level representations, show that using the combined embedding performs best compared to using either the salient or background embedding alone. Furthermore, the difference in performance between the salient and background space vary significantly between functional groups. The difference in recall between classifiers trained on background versus salient embeddings

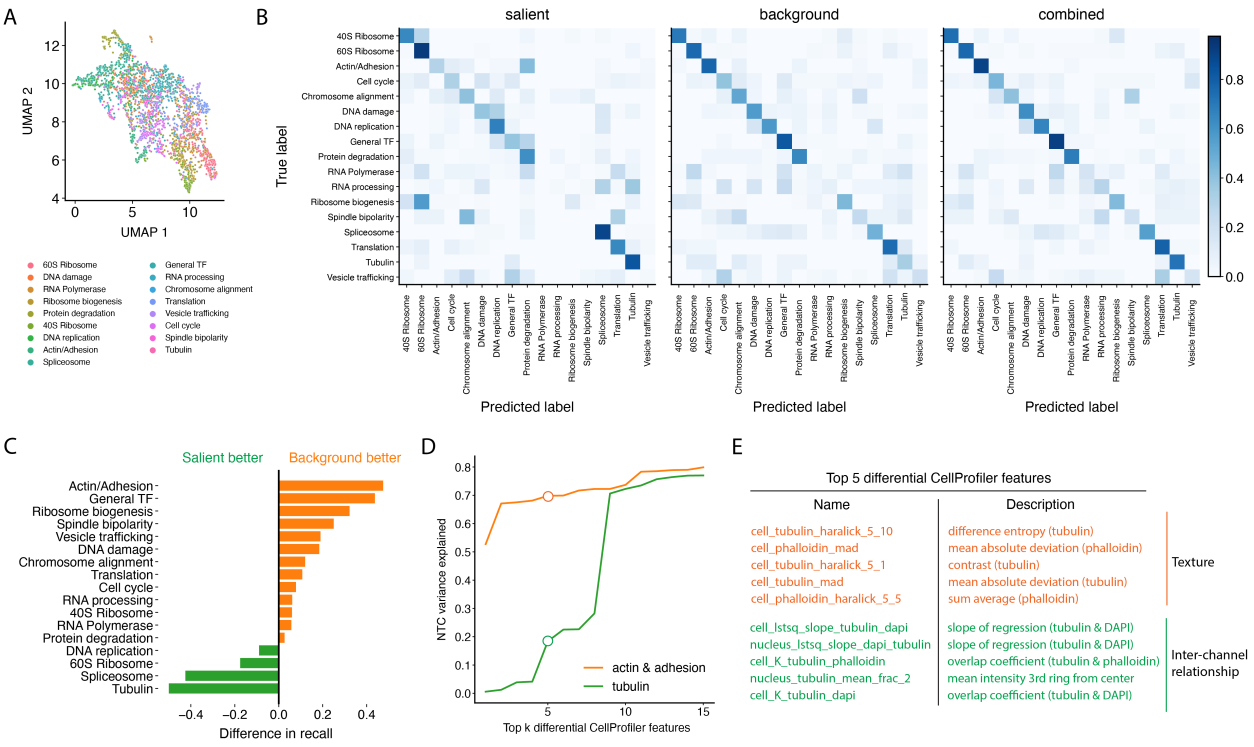


Figure 4: **The salient space better delineates perturbations that induce novel phenotypes while the background space identifies perturbations that bias the distribution of cells in existing phenotypes.** (A) UMAP projection of cell embeddings aggregated to the perturbation level, colored by the functional group of the perturbed gene. (B) Confusion matrices for separate logistic regression models trained to classify gene function from guide embeddings in different spaces, normalized row-wise. (C) Difference in recall between model trained on salient vs background embedding. (D) Percent of variations in the first PC of NTCs explained by different number of top differentially altered features, identified by multiple hypothesis testing with Bonferroni adjustment and ranked by robust z-score against NTCs. (E) Name and description of the top five differentially altered CP features.

appears in Figure 4C. The background embedding excels for predicting functional groups such as actin cytoskeleton/adhesion, spindle bipolarity, and chromosome alignment, while the salient embedding is better at predicting genes associated with tubulin and spliceosome.

We posit that certain gene functions, such as adhesion and mitosis, are accurately predicted by the background embedding because perturbing these genes likely affects phenotypes, like cell size and cell cycle stage, which naturally vary within NTCs. As a result, the altered features are predominantly captured by the background space. Conversely, perturbing genes in the tubulin group, which is only well-predicted by the salient space, produces novel features do not appear in NTCs. Indeed, using the top five altered CP features from the actin & adhesion perturbations, we can explain 70% of the variation in the NTCs' first principle component (PC), whereas only 20% can be explained by the top five altered CP features from the tubulin perturbations (Figure 4D). Comparing the actual CP features significantly altered by perturbations to either groups, we found that perturbing the actin & adhesion group primarily affects spatial heterogeneity of a molecule, while perturbing the tubulin group mainly affects cross-correlation between different molecular species (Figure 4E). This result suggests that spatial cross-correlation between molecular structures are fairly conserved, and perturbing only tubulin disrupts these cross-correlations to generate a novel phenotype rarely seen in natural cell populations.

In summary, the differences in classification performance between salient and background embedding stem from the nature of the gene perturbations, specifically whether they generate new phenotypes or merely shift the distribution of existing phenotypes, shedding light on the complex interaction between gene functions and their visual representation in the cell.

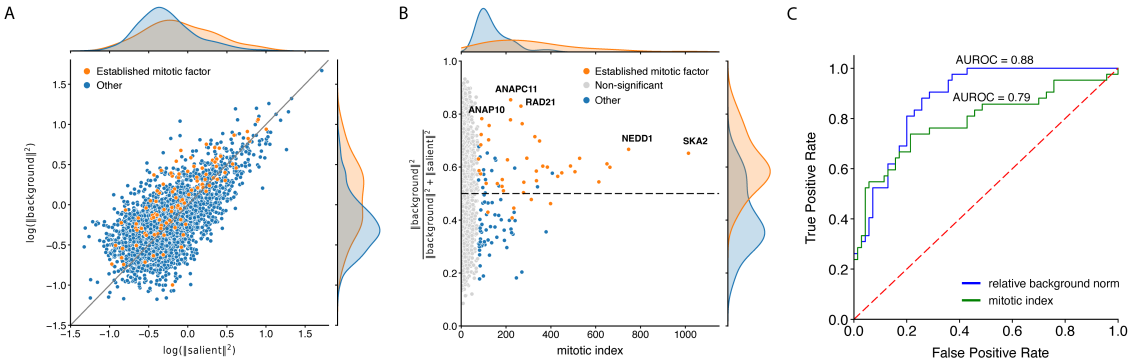


Figure 5: **Identifying mitotic factors by comparing background and salient embeddings.** (A) Logarithm of the Euclidean norm of background vs salient embeddings (normalized to have equal average across all genes) for each gene, filtered to only genes with significant difference compared to NTCs. (B) Mitotic index vs relative background norm with a select subset of established mitotic factors labeled, where significant genes (blue and orange) indicates those with a mitotic index higher than 99.9% of NTCs). (C) ROC curve for using different metrics to classify mitotic factors

3.4 Comparing Background and Salient Embeddings enables us to Identify Mitotic Factors

Cell cycle stage represents a strong source of phenotypic variation in a natural population. We can take advantage of our explicit separation of variation into background and salient spaces to identify factors that play a key role in mitosis. Namely, we found that established mitotic factors tend to have a larger norm for their background embedding compared to their salient embedding (Figure 5A), suggesting that we may use these measures to identify mitotic factors. However, genes other than mitotic factors could also have larger background embedding compared to salient. As we have already seen (Figure 4C), the background space can comprise many different kinds of biological variations beyond cell cycle. Thus to identify mitotic factors, we first filter out gene perturbations that do not produce significant changes in the proportion of mitotic cells compared to NTCs, defined as the mitotic index in Figure 5B. Cell cycle information were predicted from a support vector classifier trained on 2,500 manually annotated cells [15]. After filtering, we found that the relative background norm becomes highly predictive of whether a gene is a mitotic factor (Figure 5B). In fact, in terms of area under the ROC curve, we can achieve better classification performance for mitotic factors using the relative background norm instead of the commonly used metric of mitotic index (Figure 5C).

4 Discussion

In this work, we proposed Multi-Contrastive VAE (mcVAE), a method for disentangling perturbation effects by comparing multiple treatment groups to a single reference/control group. We applied mcVAE to a recent large-scale optical pooled screen dataset [15] consisting of over 30 millions cells spanning more than $5,000$ genetic perturbations to show that it can effectively remove technical imaging artifacts to identify perturbations that generate novel phenotypes.

Although mcVAE effectively isolated novel phenotypes in the salient space, disentanglement in the background is still a work in progress since it is made up of both technical artifacts and biological variations. We plan to extend our model to include three encoders corresponding to three latent spaces that separately captures technical noise, natural phenotypic variations, and novel perturbation-induced phenotypes. We can incorporate kernel-based independence measures [19] to facilitate the enforcement of independence statements between the technical noise latent variables and the perturbation label.

Exploring deeper neural network architectures is another important extension of this work. In this current work, we used a relatively simple encoder/decoder architecture with only five convolutional/deconvolutional layers. A deeper architecture might foster a finer granularity in the detection of subtle phenotypic patterns which are otherwise overshadowed in shallow architectures. Furthermore, we used a small number of dimensions for the salient and background space (32 dimensions each). Increasing the number of latent dimensions in our mcVAE model can potentially enhance the representation of complex, high-dimensional data, allowing for a more nuanced understanding of genetic perturbations.

References

[1] Christoph Bock, Paul Datlinger, Florence Chardon, Matthew A Coelho, Matthew B Dong, Keith A Lawson, Tian Lu, Laetitia Maroc, Thomas M Norman, Bicna Song, et al. High-content CRISPR screening. *Nature Reviews Methods Primers*, 2(1):8, 2022.

[2] David Feldman, Avtar Singh, Jonathan L Schmid-Burgk, Rebecca J Carlson, Anja Mezger, Anthony J Garrity, Feng Zhang, and Paul C Blainey. Optical pooled screens in human cells. *Cell*, 179(3):787–799, 2019.

[3] Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, et al. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9):849–863, 2017.

[4] 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.

[5] Erick Moen, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. Deep learning for cellular image analysis. *Nature Methods*, 16(12):1233–1246, 2019.

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

[8] James Y Zou, Daniel J Hsu, David C Parkes, and Ryan P Adams. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[9] Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):2134, 2018.

[10] Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019.

[11] Adrià Ruiz, Oriol Martinez, Xavier Binefa, and Jakob Verbeek. Learning disentangled representations with reference-based variational autoencoders. *arXiv preprint arXiv:1901.08534*, 2019.

[12] Kristen A Severson, Soumya Ghosh, and Kenney Ng. Unsupervised learning with contrastive latent variable models. In *AAAI Conference on Artificial Intelligence*, pages 4862–4869, 2019.

[13] Ethan Weinberger, Nicasia Beebe-Wang, and Su-In Lee. Moment matching deep contrastive latent variable models. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[14] Ethan Weinberger, Romain Lopez, Jan-Christian Huetter, and Aviv Regev. Disentangling shared and group-specific variations in single-cell transcriptomics data with multiGroupVI. In *Machine Learning in Computational Biology*, volume 200 of *Proceedings of Machine Learning Research*, pages 16–32, 21–22 Nov 2022.

[15] Luke Funk, Kuan-Chung Su, Jimmy Ly, David Feldman, Avtar Singh, Brittania Moodie, Paul C Blainey, and Iain M Cheeseman. The phenotypic landscape of essential human genes. *Cell*, 185(24):4634–4653, 2022.

[16] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[17] Safiye Celik, Jan-Christian Huetter, Sandra Melo-Carlos, Nathan Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta Fay, Berton Earnshaw, and Imran Haque. Biological cartography: Building and benchmarking representations of life. In *NeurIPS Workshop on Learning Meaningful Representations of Life*, 2022.

[18] George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic acids research*, 51(D1):D539–D545, 2023.

[19] Romain Lopez, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. Information constraints on auto-encoding variational Bayes. *Advances in Neural Information Processing Systems*, 31, 2018.

[20] Luke Funk. Single-cell data for "The phenotypic landscape of essential human genes", 2022. URL `https://doi.org/10.7910/DVN/VYKTI5`.

[21] Luke Funk. The phenotypic landscape of essential human genes. `https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD394`, 2022.

A Supplementary Material

A.1 Data acquisition and preprocessing

CellProfiler features with corresponding metadata, including cell cycle stage and perturbation label (gene targeted by CRISPR), for each cell were obtained directly from the online repository Harvard Dataverse [20]. These features were already batch-corrected by standardizing against the NTCs in the corresponding batch.

Raw microscopy images each covering a large field of view with many cells were downloaded from BioImage Archive [21], followed by imaging channel alignment using phase cross-correlation. For each raw image, pixels with intensity values in the top and bottom 0.1% were clipped. Finally we obtain individual cell patches by using the cell positional values from CellProfiler data to obtain a 100 pixel by 100 pixel bounding box around each cell which we use to represent individual cells for model training, this also led us to drop cells within 50 pixel of the tile edge.

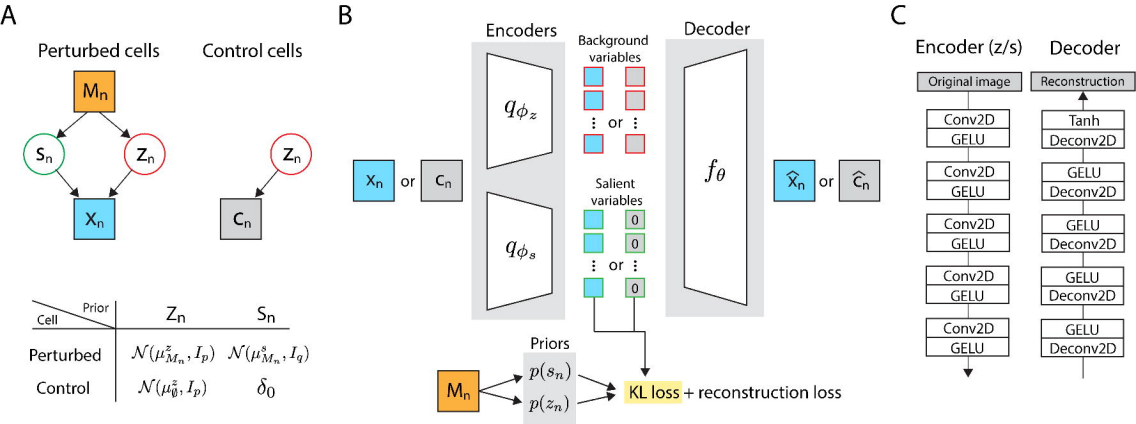
Acknowledgments and Disclosure of Funding

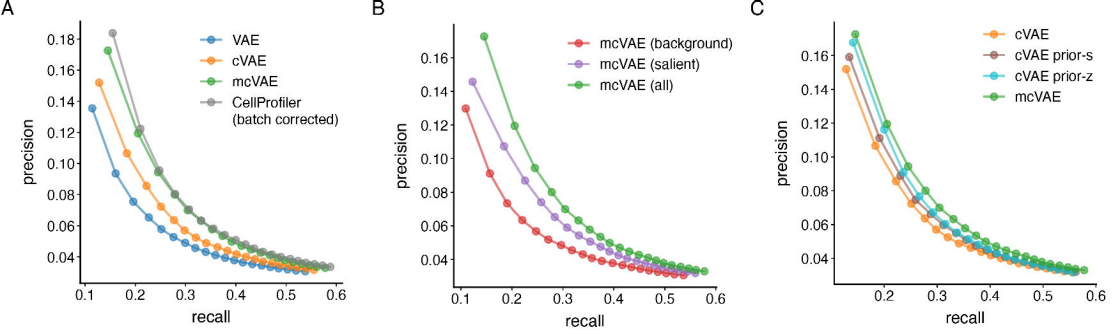
We thank Avtar Singh, David Richmond, Mahtab Bigverdi, Anqi Zhu, Rebecca Boiarsky, Xinming Tu, and Kexin Huang for insightful feedback throughout the duration of this project which greatly improved this work. We also thank members of the Regev Lab and the Artificial Intelligence/Machine Learning department at Genentech for providing constructive feedback on the results presented in this work.

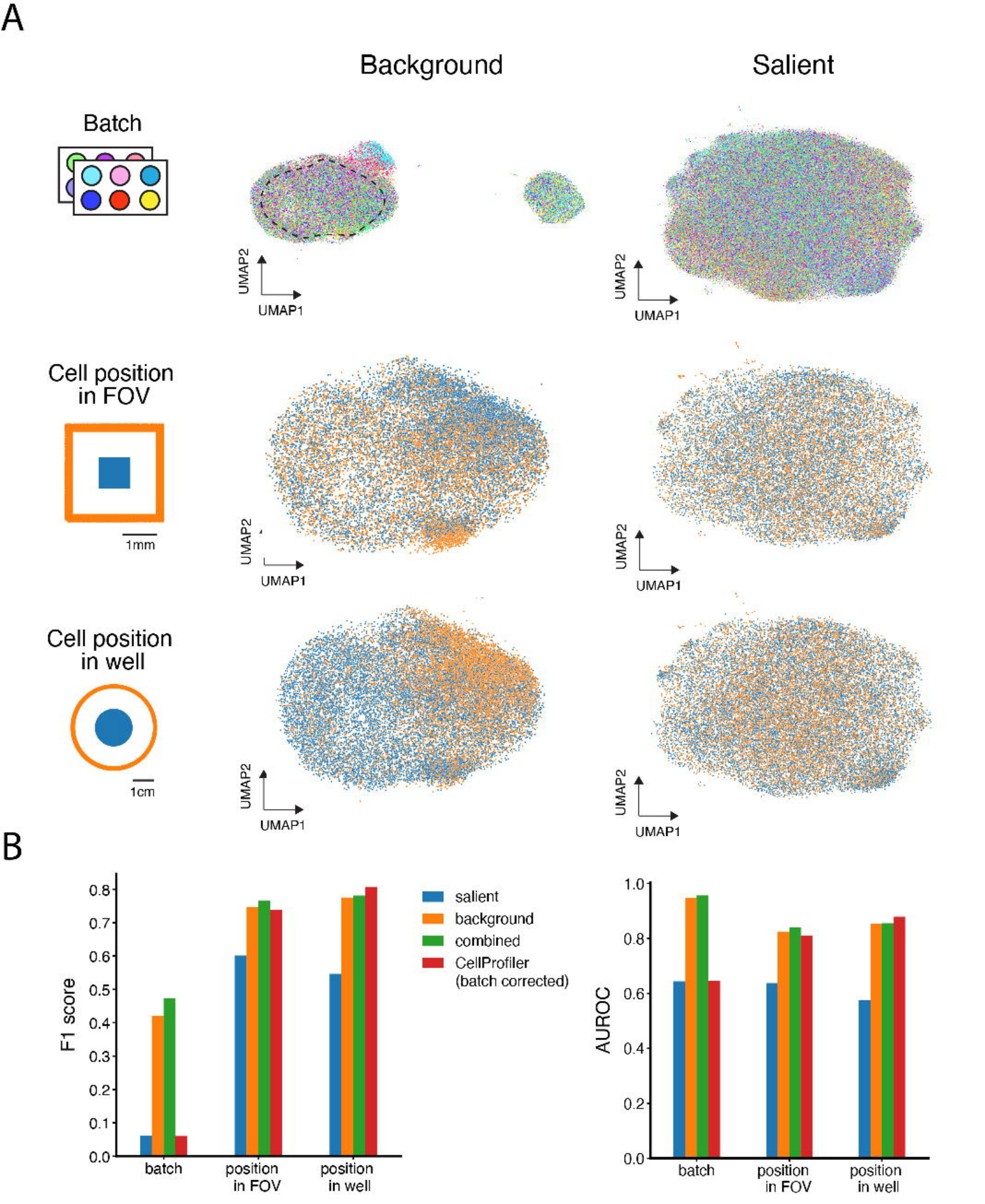
Disclosures: This work was performed while Zitong Jerry Wang was employed as an intern at Genentech. Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Heming Yao, Philipp Haslovsky and Burkhard Hoeckendorf are employees of Genentech, and Jan-Christian Hütter, Heming Yao, Philipp Haslovsky and Burkhard Hoeckendorf have equity in Roche. Aviv Regev is a co-founder and equity holder of Celsius Therapeutics and an equity holder in Immunitas. She was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics, and Asimov until July 31st, 2020; she has been an employee of Genentech since August 1st, 2020, and has equity in Roche.

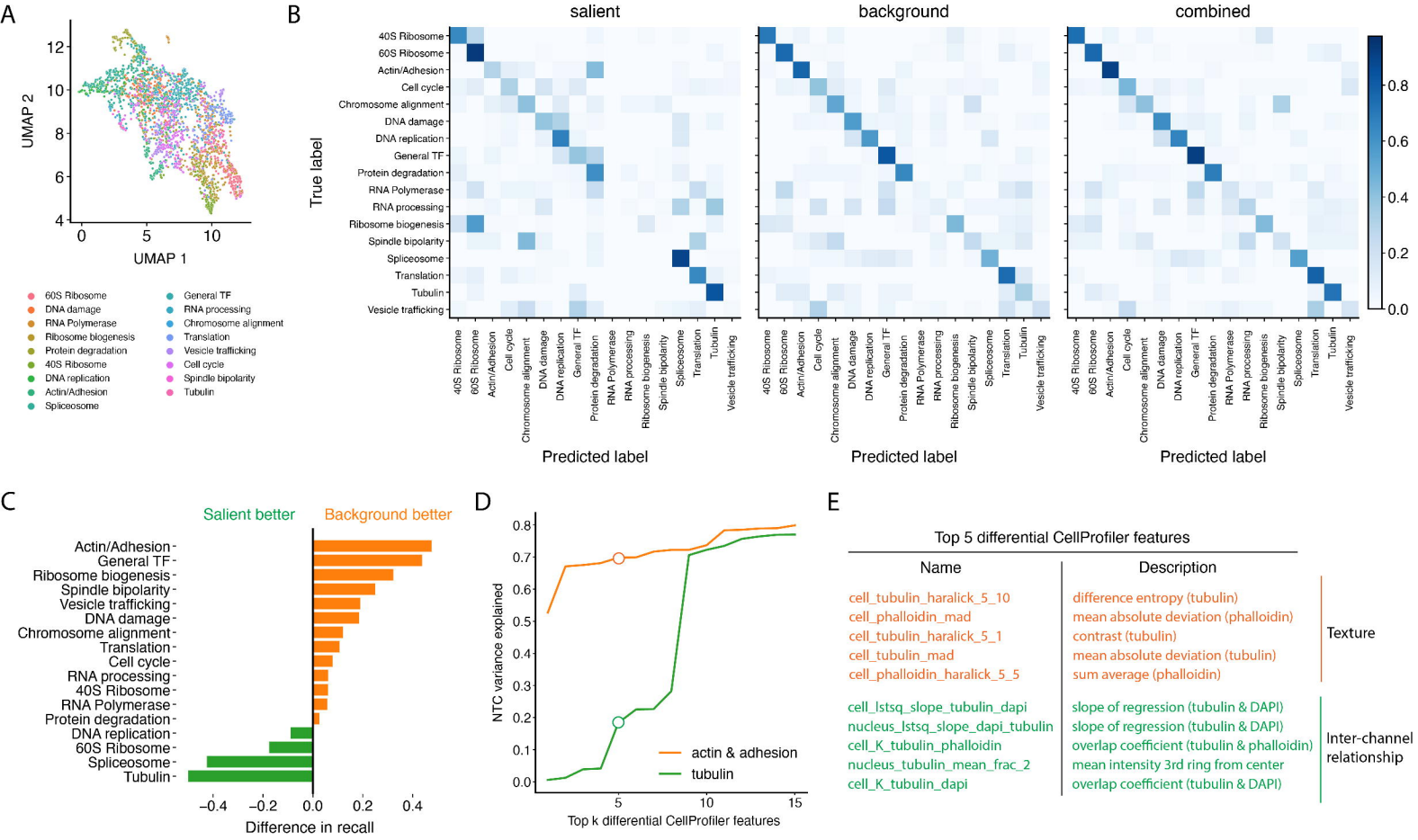
Code Availability Statement

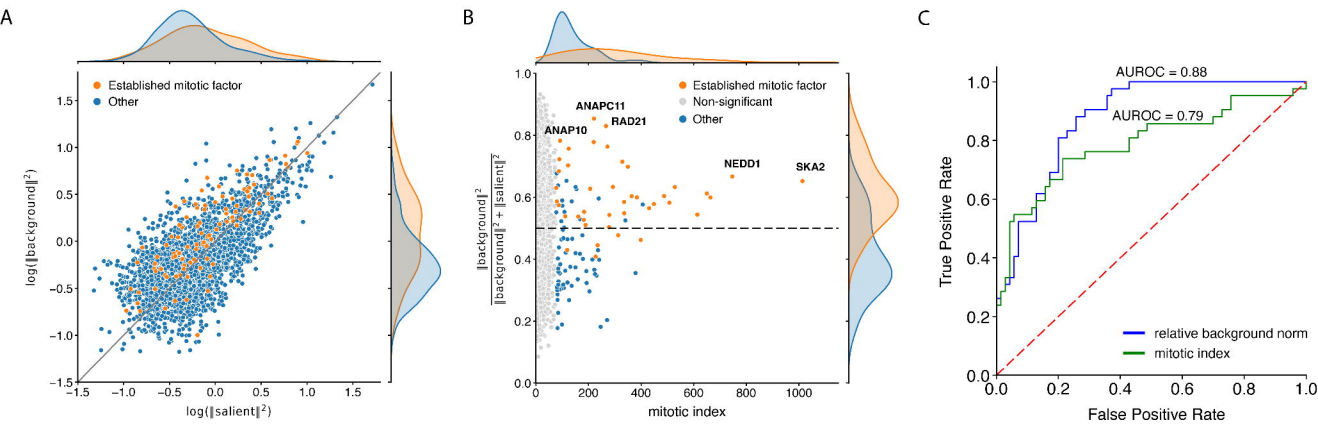
We implemented our Multi-Contrastive VAE model in PyTorch, with all implementation code for model training, analysis, and figure generation available at `https://github.com/Genentech/contrastive-ops`.


A
Perturbed cells
Control cells
Mn
Sn
Zn
Xn
Zn
Cn
Prior
Cell
Zn
Sn
Perturbed
$\mathcal{N}(\mu^z_{M_n}, I_p)$
$\mathcal{N}(\mu^s_{M_n}, I_q)$
Control
$\mathcal{N}(\mu^z_{\emptyset}, I_p)$
δ_0
B
Encoders
Background variables
Decoder
q_{ϕ_z}
q_{ϕ_s}
Salient variables
f_θ
Xn or Cn
or
0
$\hat{x}_n$ or $\hat{c}_n$
Priors
Mn
$p(s_n)$
$p(z_n)$
KL loss + reconstruction loss
C
Encoder (z/s)
Decoder
Original image
Reconstruction
Conv2D
GELU
Tanh
Deconv2D


A
precision
recall
VAE
cVAE
mcVAE
CellProfiler
(batch corrected)
B
precision
recall
mcVAE (background)
mcVAE (salient)
mcVAE (all)
C
precision
recall
cVAE
cVAE prior-s
cVAE prior-z
mcVAE


A
Background
Salient
Batch
UMAP2
UMAP1
Cell position in FOV
1mm
Cell position in well
1cm
B
F1 score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
batch
position in FOV
position in well
salient
background
combined
CellProfiler (batch corrected)
AUROC
0.0
0.2
0.4
0.6
0.8
1.0


A
UMAP 2
UMAP 1
60S Ribosome
DNA damage
RNA Polymerase
Ribosome biogenesis
Protein degradation
40S Ribosome
DNA replication
Actin/Adhesion
Spliceosome
General TF
RNA processing
Chromosome alignment
Translation
Vesicle trafficking
Cell cycle
Spindle bipolarity
Tubulin
B
salient
background
combined
True label
Predicted label
C
Salient better
Background better
Difference in recall
D
NTC variance explained
Top k differential CellProfiler features
actin & adhesion
tubulin
E
Top 5 differential CellProfiler features
Name
Description
cell_tubulin_haralick_5_10
difference entropy (tubulin)
cell_phalloidin_mad
mean absolute deviation (phalloidin)
cell_tubulin_haralick_5_1
contrast (tubulin)
cell_tubulin_mad
mean absolute deviation (tubulin)
cell_phalloidin_haralick_5_5
sum average (phalloidin)
Texture
cell_lstsq_slope_tubulin_dapi
slope of regression (tubulin & DAPI)
nucleus_lstsq_slope_dapi_tubulin
slope of regression (tubulin & DAPI)
cell_K_tubulin_phalloidin
overlap coefficient (tubulin & phalloidin)
nucleus_tubulin_mean_frac_2
mean intensity 3rd ring from center
cell_K_tubulin_dapi
overlap coefficient (tubulin & DAPI)
Inter-channel relationship


A
Established mitotic factor
Other
log(‖background‖²)
log(‖salient‖²)
B
Established mitotic factor
Non-significant
Other
ANAPC11
RAD21
ANAP10
NEDD1
SKA2
‖background‖² / (‖background‖² + ‖salient‖²)
mitotic index
C
AUROC = 0.88
AUROC = 0.79
relative background norm
mitotic index
True Positive Rate
False Positive Rate