Article

Patterns

Generative modeling of single-cell gene expression for dose-dependent chemical perturbations

Highlights

- Predicts chemical perturbations in gene expression across cell types
- Predicts response to multiple doses of a chemical
- Enables biological interpretation of model predictions
- "Pseudo-dose" metric evaluates cell-specific chemical sensitivity

Authors

Omar Kana, Rance Nault, David Filipovic, Daniel Marri, Tim Zacharewski, Sudin Bhattacharya

Correspondence

sbhattac@msu.edu

In brief

Variational autoencoders can predict chemical perturbations across cell types using vector arithmetic. However, vector arithmetic alone cannot predict perturbations in single-cell gene expression accurately in animal studies across multiple doses. We utilize a regression-based method to improve on *in vivo* predictions by accounting for cell-type-specific differences in gene expression response. We then extend this model to predict the response to multiple doses of a chemical and derive a metric to characterize chemical sensitivity in individual cells.

Kana et al., 2023, Patterns *4*, 100817
August 11, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.patter.2023.100817

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Generative modeling of single-cell gene expression for dose-dependent chemical perturbations

Omar Kana,[1,2,3] Rance Nault,[2,4] David Filipovic,[3,5,6] Daniel Marri,[3,5] Tim Zacharewski,[2,4] and Sudin Bhattacharya[1,2,3,5,7,8,*]
[1]Department of Pharmacology and Toxicology, Michigan State University, East Lansing, MI 48824, USA
[2]Institute for Integrative Toxicology, Michigan State University, East Lansing, MI 48824, USA
[3]Institute for Quantitative Health Science & Engineering, Michigan State University, East Lansing, MI 48824, USA
[4]Department of Biochemistry and Molecular Biology Michigan State University, Michigan State University, East Lansing, MI 48824, USA
[5]Department of Biomedical Engineering, Michigan State University, East Lansing, MI 48824, USA
[6]Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
[7]Senior author
[8]Lead contact
*Correspondence: sbhattac@msu.edu
https://doi.org/10.1016/j.patter.2023.100817

THE BIGGER PICTURE Cellular response to chemical perturbation is highly heterogeneous and dose dependent. It would be impossible to experimentally characterize the risks of chemical or drug exposure across all relevant combinations of cell types, chemicals, and doses. We introduce scVIDR, a computational method that utilizes recent advances in generative deep learning to address this challenge. Across a range of chemical exposure scenarios, we show that after training on available single-cell gene expression data, scVIDR can predict perturbations across untested cell types and doses. We envision that scVIDR will help reduce the need for repeated animal testing across tissues, chemicals, and doses.

1 **2** 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Single-cell sequencing reveals the heterogeneity of cellular response to chemical perturbations. However, testing all relevant combinations of cell types, chemicals, and doses is a daunting task. A deep generative learning formalism called variational autoencoders (VAEs) has been effective in predicting single-cell gene expression perturbations for single doses. Here, we introduce single-cell variational inference of dose-response (scVIDR), a VAE-based model that predicts both single-dose and multiple-dose cellular responses better than existing models. We show that scVIDR can predict dose-dependent gene expression across mouse hepatocytes, human blood cells, and cancer cell lines. We biologically interpret the latent space of scVIDR using a regression model and use scVIDR to order individual cells based on their sensitivity to chemical perturbation by assigning each cell a "pseudo-dose" value. We envision that scVIDR can help reduce the need for repeated animal testing across tissues, chemicals, and doses.

INTRODUCTION

In 2010, Sydney Brenner suggested that it is possible to deduce the physiology of biological systems by understanding the interactions and behaviors of their constituent units.[1] The appropriate unit, in his opinion, was the cell. Single-cell sequencing (scSeq) has revolutionized the study of cell biology. With the ability to capture the transcriptomic state of thousands of cells at once, a fine-grained picture of the organization of cell physiology has begun to emerge.[2] Much of the effort in scSeq has been made in the realm of cell-type/-state discovery,[3,4] cellular development,[5–8] and disease progression.[9,10] These represent natural applications of scSeq, especially regarding the spatial and temporal dynamics of cellular systems and their interactions. However, relatively little attention has been given to how cells respond to environmental signals like chemical exposures, which in addition to being spatial and temporal are also chemical and dose dependent.

Broadly, cells exhibit the ability to recognize and respond to external stimuli. This process is mediated by a coordinated set

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

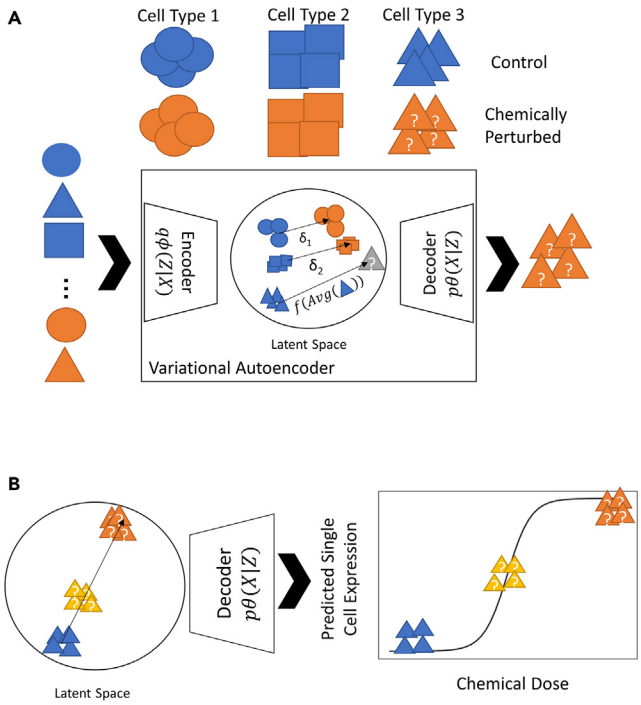


Figure 1. Schematic of scVIDR for prediction of response to single and multiple doses for some unknown cell type

(A) Outline of the scVIDR model for expression prediction for unknown single-dose response in cell type 3. Training is done using cell types 1 and 2 as input to a variational autoencoder model. The difference between the centroids of latent representations of the control and treated groups, δ_1 and δ_2, are used as input into a linear regression model. The linear regression model is then used to predict the δ_3 of the test cell type 3. We then use the decoder portion of the model to convert the latent space predictions back into gene expression space.

(B) Use of scVIDR for prediction of the unknown response of multiple doses for cell type 3. Log-linear interpolation on δ_3 is used to predict dose-dependent changes in gene expression in the latent space. The latent space representations are then projected back into gene expression space using the decoder.

of extracellular and intracellular interactions that transduce resulting signals into cellular responses.[11] These responses, as a function of dose, define dose-response curves.[12] The dose-response curve is heavily dependent on the type of cell and its internal state.[13,14] Thus, even cells of the same type can respond to the same exposure in a heterogeneous manner.[15] scSeq provides a comprehensive measure of the transcriptome of a cell and captures the inherent variation among cells of the same type. This makes scSeq a useful tool in the study of chemical perturbations of biological systems.

However, a comprehensive cell atlas of chemical perturbations is impossible to assemble given the vast number of combinations of dose, exposure duration, and cell types.[16] Recently developed resources like scPerturb[17] and the multiplexed interrogation of gene expression through single-cell RNA sequencing (MIX-seq) protocol[18] cover a meaningful but relatively small portion of this space. Algorithms that generalize chemical perturbations across cell state and dose can provide better estimates of the cartography of the chemical perturbation space. In this work, we use deep generative modeling to computationally predict cellular response across dose and cell types. We use a class of deep neural networks for dimensionality reduction called autoencoders. Specifically, we use a variational autoencoder[19] (VAE), which relies on Bayesian priors to encode single-cell data into a latent distribution. VAEs have been used to model several technical aspects unique to single-cell data, including statistical confounders such as library size and batch effects[20] and zero inflation.[21]

In perturbational single-cell biology, autoencoder models such as scGen22 have been able to predict the response of interferon β (IFN-β)-treated peripheral blood mononuclear cells (PBMCs). However, for considering more complicated *in vivo* perturbations, existing models do not consider cell-type-specific effects in predicting the mean expression of differentially expressed genes (DEGs). Advances in other autoencoder frameworks such as the compositional perturbational autoencoder (CPA)[16] aim to deal with these issues by trying to infer basal state from the data by modeling covariates with different autoencoders and then iteratively composing them when performing predictions for a particular set of conditions. While promising, CPA can only work with vary large data samples (relative to other perturbational autoencoders), as the model needs to learn a latent space for each covariate. Thus, for confident prediction, CPA will need datasets that already have a great deal of the perturbational space mapped. Additionally, most perturbational autoencoder frameworks are uninterpretable in terms of the quantitative relationship between latent space and expression prediction. Thus, it is difficult to ascertain which specific genes the model uses to predict differential gene expression after treatment. Thus, there is a need for simpler models that better account for the complexity of *in vivo* experiments, that predict high doses from less data, and that provide more informative interpretations at the level of individual genes.

Here, we propose single-cell variational inference of dose-response (scVIDR), which builds on latent space vector arithmetic when using VAEs to study single-cell perturbations (Figure 1). scVIDR predicts cell-type-specific DEG expression and approximates high-dose experiments better than other state-of-the-art algorithms. We also use scVIDR to interpret the latent space using linear models to assess the pathways involved in the single-cell dose-response. We accomplish this across several datasets including the dose-response of liver cells to 2,3,7,8 tetrachlorodibenzo-*p*-dioxin (TCDD) *in vivo*,[22,23] PBMCs treated with IFN-β,[24] and a multiplexed dataset of 188 different drug combinations applied to three prominent cancer cell lines (sci-Plex[25]).

We use data from a single-nucleus dose-response experiment in livers from mice gavaged with TCDD as a case study for *in vivo* dose-response prediction.[22,23] Hepatic responses to TCDD represent an interesting case study, as its canonical receptor, the aryl hydrocarbon receptor (AhR), is unevenly expressed along the hepatic lobule, the functional unit of the liver. AhR is more highly expressed in the centrilobular region compared with the portal region (Figure S1).[26] Thus, not only does response to TCDD vary across different cell types in the liver, but it also varies within cell types (such as hepatocytes) along the portal to the central axis of the liver lobule.[22,27] To model response variation between cell types, the latent space of the VAE is used to order hepatocytes with respect to their transcriptomic response to TCDD and thus align all hepatocytes along a "pseudo-dose" axis.

RESULTS

scVIDR predicts single-dose, single-cell perturbation expression better than other state-of-the-art algorithms

According to the manifold hypothesis, high-dimensional data often lay on a lower-dimensional, latent manifold.[28] For single-cell data, this is a reasonable assumption given that the expression of one gene is often highly dependent on the expression of other genes encoding transcription factors and is functionally constrained by the process of evolution.[29] Further evidence of this can be seen in the extensive use and success of dimensionality reduction algorithms in the analysis of scSeq data.[30] Lower-dimensional representations of single-cell data are at the heart of many single-cell gene expression analysis methods such as trajectory inference.[31] One method of interest is modeling of the latent manifold using neural networks. These latent manifolds have been shown to simplify complex relationships in single-cell gene expression data.[32–34] Specifically, simple vector arithmetic on such spaces can predict *in vitro* chemical perturbations with high accuracy.[16,35] However, the accuracy of such models when predicting *in vivo* dose-responses is inconsistent.

We begin by considering a single-cell gene expression dataset $X = \{x_i\}_{i=1}^{N}$ consisting of N cells, where x_i represents the expression profile of cell i. We assume that gene expression is generated by some continuous random process involving a lower-dimensional random variable z. The generative process that describes the mapping from z to X is given by the probability distribution, $p_\theta(X|z)$. Thus, given that we know X and not z, we would like to approximate the probability distribution that maps X to z, $p_\theta(z|X)$. Since calculating $p_\theta(z|X)$ is usually intractable, we use a neural network, the encoder, to approximate it using a different Gaussian distribution, $q_\varphi(z|X)$. To map values back from z to X, we use a second neural network, the decoder, to approximate $p_\theta(X|z)$. In practice, both the encoder and decoder are trained together to minimize the reconstruction error of the decoder and the difference between the prior distribution and the encoder distribution.

We initially developed models for a single-dose chemical perturbation where we characterize whether a cell has been treated with a set concentration of the chemical of interest with the indicator variable t (Figure 1A). We set $t = 1$ for cells that have been treated with the chemical (treatment) and $t = 0$ for cells that have not been treated (control). Our dataset contains c cell types within both the $t = 0$ and $t = 1$ groups. Each time a model is evaluated, one treated cell type is withheld from training and used in evaluation. In standard VAE vector arithmetic (scGen), the latent space representation of the perturbation of some cell type A is approximated by $\widehat{z}_{i,A,t=1} = z_{i,A,t=0} + \delta$. $z_{i,A,t}$ is the latent gene expression representations of cell type A,[35] and δ is the difference between the centroids of the treated and control training groups in the latent space. When we compare the difference of centroids between the treated and control groups, δ_c, of individual cell types with δ, we see that cell-type-specific differences vary greatly in a principal-component analysis (PCA) projection (Figure S2A). Examination of the magnitudes (Figure S2B) and the directions of each cell's δ_c (Figure S2C) in high-dimensional space show that δ_c diverges greatly from δ. Hence, we calculate $\widehat{\delta}_{c=A}$, a function of the mean latent representation of the control group of cell type A. We approximate this function by training a linear regression model with the other cell types on the latent space (experimental procedures) and show that $\widehat{\delta}_{c=A}$ better matches the ground truth $\delta_{c=A}$ (Figure S2). It should be noted that when there is only one cell type available for training, for all practical purposes, scVIDR is equivalent to scGen (Figure S7).

We applied this model to the case of a single dose of TCDD administered to mice. Gene expression was measured with single-nucleus RNA-seq (snRNA-seq) originating from the mouse liver. We set $t = 0$ for unperturbed gene expression and $t = 1$ for gene expression perturbed by 30 μg/kg TCDD. The dataset covered 6 different liver cell types: cholangiocytes, endothelial cells, stellate cells, central hepatocytes, portal hepatocytes, and portal fibroblasts (Figure 2). Our training set (Figure 2A) consisted of all control and TCDD-treated cell types except for TCDD-treated portal hepatocytes, which were used for model evaluation. We compared the performance of scGen, scPreGAN,[36] CellOT,[37] and scVIDR (our method) on the top 5,000 highly variable genes (HVGs) and the top 100 DEGs. When predicting the gene expression of portal hepatocytes, each method generated a set of virtual portal hepatocytes (Figure 2B). We then computed the average expression of each gene across all cells and compared the average gene expression in predicted cells versus cells derived from snRNA-seq experiments. Across HVGs, the scVIDR model yielded an average R^2 of 0.92 (Figure 2C). Across DEGs, scVIDR produced an average R^2 of 0.81 (Figure 2C). Continuing the evaluation across all cell types (Figure 2D), leaving out one cell-type perturbation at a time as described above for portal hepatocytes, our model outperformed all other models (with $p < 0.001$, one sided Mann-Whitney U test) when evaluated on both HVGs and DEGs.

We had similar results for IFN-β-treated PBMCs (Figure S3).[24] Here, $t = 1$ for PBMCs treated with IFN-β, and $t = 0$ for untreated PBMCs (Figure S3A). Across HVGs, the models yielded R^2 values of 0.97, 0.92, 0.77, and 0.66, and across DEGs, they yielded R^2s of 0.96, 0.86, 0.80, and 0.84 for scVIDR, scGen, scPreGAN, and CellOT, respectively (Figure S3C). When accuracy was assessed for all cell types, scVIDR significantly outperformed all other models (Figure S3D).

To test if scVIDR can perform out-of-distribution predictions robust to experimental batch effects and diverse genetic backgrounds, we test scVIDR on two additional experiments. In the first experiment, we recapitulate results from Lotfollahi et al.,[35] in which we predict perturbations across studies (in this case, we look at IFN-β perturbation of PBMCs from Kang et al.[24] and try to predict it in PBMCs from Zheng et al.[38]). We show that scVIDR can predict biologically plausible perturbations across studies (Figure S8). In the second experiment, we show that scVIDR can better predict LPS6 perturbation in rats ($R^2 = 0.92$ for HVGs) using perturbations from other species (pig, rabbit, and mouse)[39] than scGen ($R^2 = 0.91$ for HVGs), scPreGAN ($R^2 = 0.63$ for HVGs), and CellOT ($R^2 = 0.23$ for HVGs) (Figure S9). In both experiments, we show that scVIDR can be used to predict perturbations not only across cell types but also across multiple perturbation studies and models.

scVIDR accurately predicts the transcriptomic response for multiple doses across cell types

Next, we predicted the response for multiple doses of TCDD (Figure 1B). Here, p is equal to the magnitude of the perturbation,

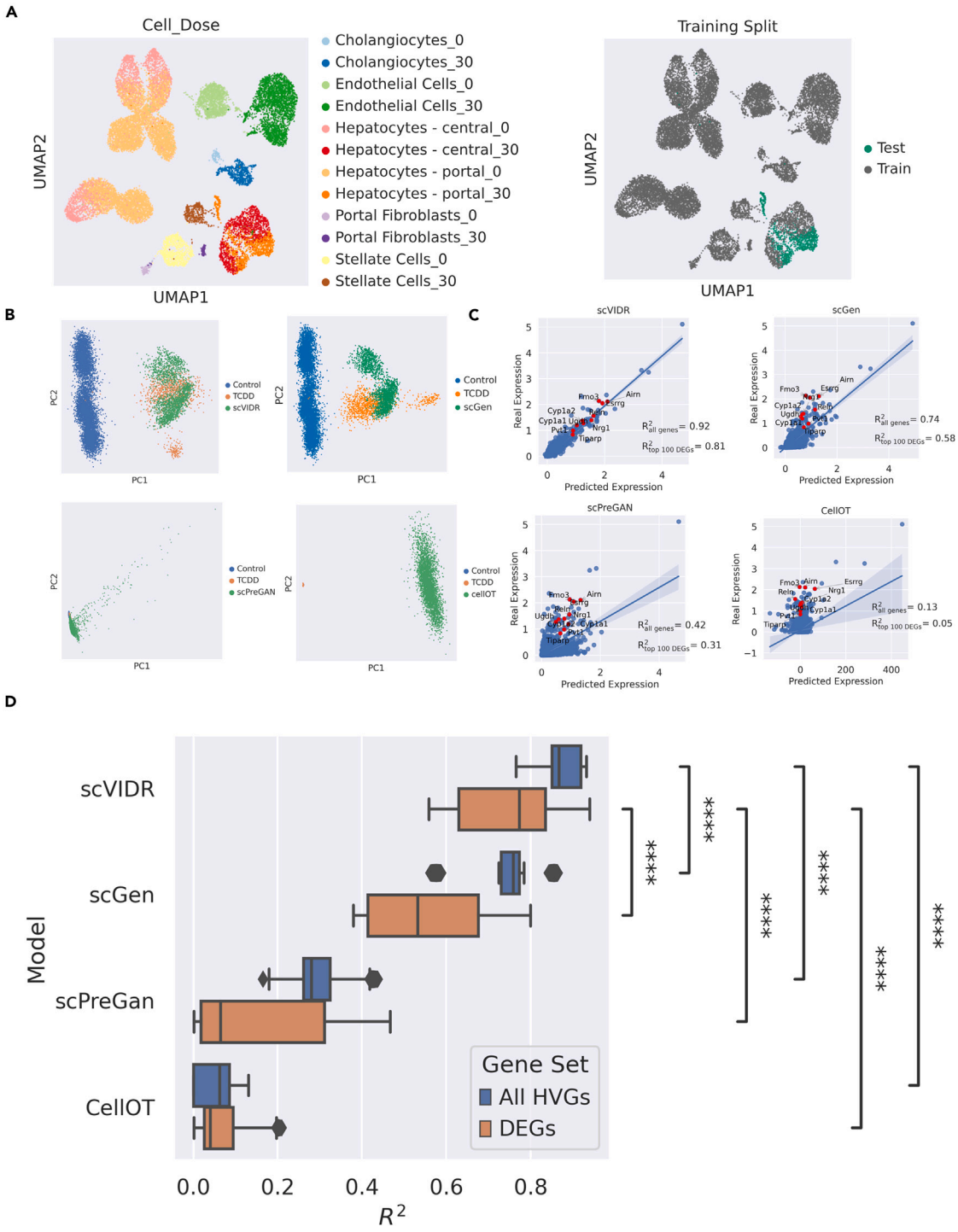


Figure 2. Prediction of *in vivo* single-cell gene expression of portal hepatocytes from mice treated with 30 μg/kg TCDD

(A) Uniform manifold approximation and projection (UMAP) of the latent space representation of control and treated single-cell gene expression. Each cell type and dose in μg/kg combination and by the train-test split for model training is represented by different colors. In the example in the figure, TCDD-treated portal hepatocytes were used as a test set.

(B) PCA plots of predicted portal hepatocyte responses following treatment with 30 μg/kg TCDD using scGen, scVIDR, *scPreGAN*, and *CellOT*.

(C) Regression plots of each model. Each point represents the mean expression of a particular gene. Red points represent the top ten differentially expressed genes. Shaded region around regression line represents the 95% confidence interval.

(D) Boxplot of R^2 values for predictions across all liver cell types treated with 30 μg/kg TCDD. Calculation of the mean R^2 across all highly variable genes (blue). Calculation of the mean R^2 across the top 100 differentially expressed highly variable genes (orange). Prediction performance distributions were compared using a one-sided Mann-Whitney U test. ****$p \leq 0.0001$.

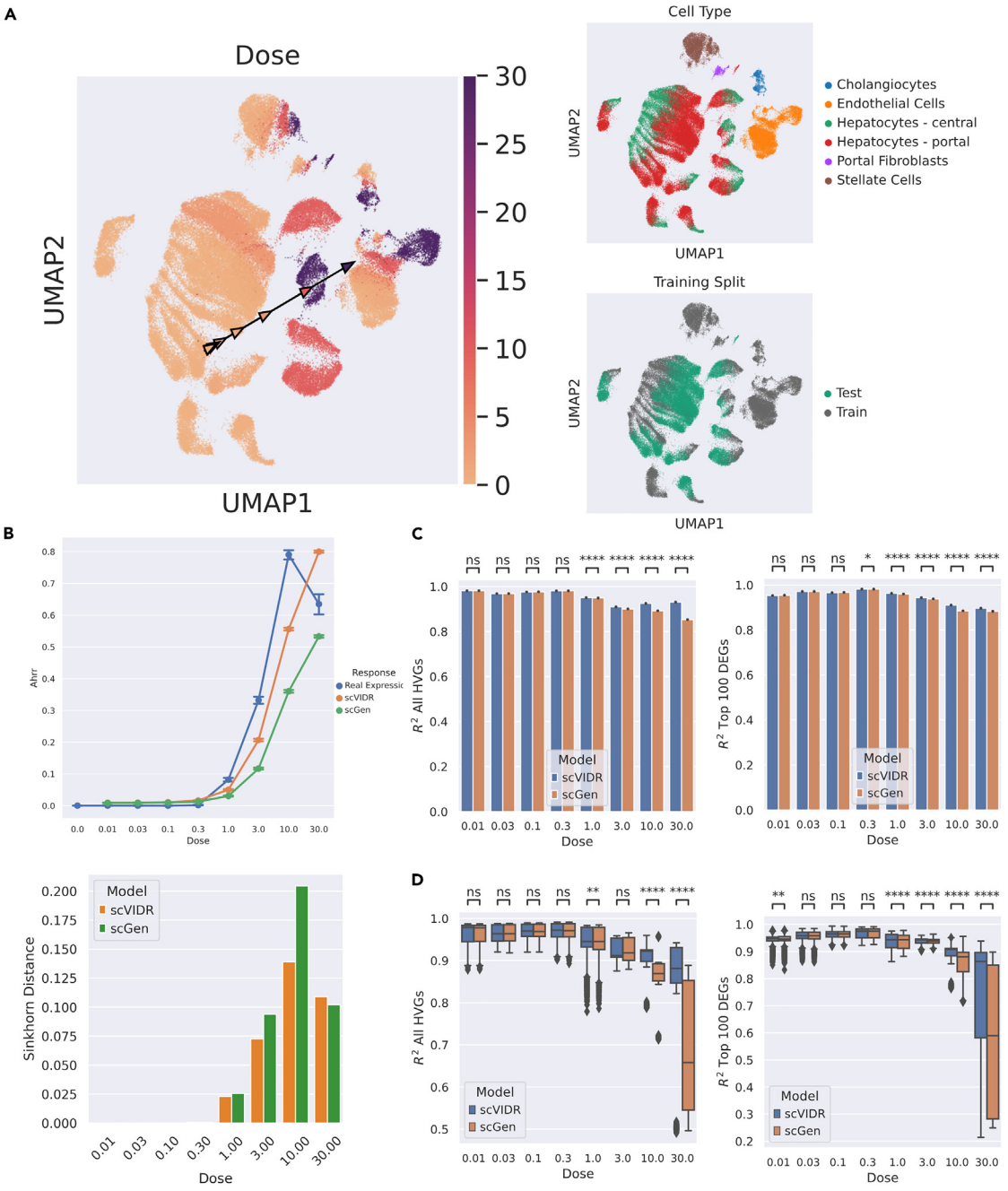


Figure 3. Prediction of *in vivo* single-cell TCDD dose-response across cell types from mouse liver

(A) UMAP of the latent space representation of single-cell gene expression across TCDD dose-response. Cells are colored by dose (μg/kg), cell type, and test-training split. Arrows on UMAP represent a δ calculated on UMAP space, with each arrowhead representing a specific dose denoted by its color.

(B) Dose-response prediction for the *Ahrr* gene using scVIDR and scGen. The differences between the predicted and true distributions of *Ahrr* at each dose are measured via the Sinkhorn distance. Bars represent standard error of expression.

(C) Bar plots of the R^2 scores of the gene expression means in portal hepatocytes for all highly variable genes and the top 100 differentially expressed genes. Significance was determined by the one-sided Mann-Whitney U test. *p between 0.05 and 0.01; ****p $\leq$ 0.0001.

(D) Boxplot of the distribution of R^2 scores across all cell types in liver tissue. **p between 0.01 and 0.001.

which in our case is equivalent to the dose. Thus, $p = 0$ represents expression at dose 0, and $p = 30$ represents expression at dose 30, where the dose is in units of μg/kg in Figure 3 and of nM in Figure S4. As with the single-dose case, we train the model on the dose-response data for all cell types except one, for which only the $p = 0$ condition is kept. We calculate the $\hat{\delta}_c$ (experimental procedures; Figure 3A), which is the estimated difference of means between the highest dose and the untreated

groups. For scVIDR, intermediate doses are then calculated on the latent space by interpolating log linearly on the $\hat{\delta}_c$. For scGen, we log linearly interpolate on δ (experimental procedures). Finally, those latent space representations are decoded back into gene expression space using the decoder portion of each of the models.

We analyzed a mouse liver snRNA-seq dataset that included 8 doses (p = [0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10, 30]) of TCDD and a control (p = 0) in μg/kg (Figure 3). scVIDR outperforms scGen in approximating expression across the dose-response of TCDD in mouse liver. We used the mean R^2 score across all evaluated genes as our performance metric (Figure 3B). scVIDR significantly outperformed scGen at predicting HVGs and DEGs for doses >0.3 μg/kg (Mann-Whitney one-sided U test $p < 0.001$). scVIDR predicts the important TCDD receptor repressor gene, *Ahrr*, at doses 1, 3, and 10 μg/kg in portal hepatocytes better than scGen (Figure 3C). When predicting all other cell types (cholangiocytes, endothelial cells, stellate cells, central hepatocytes, portal hepatocytes, and portal fibroblasts), scVIDR significantly outperformed scGen only at the highest doses of 10 and 30 μg/kg on prediction of all HVGs (Figure 3D). When predicting on just the DEGs, scVIDR significantly outperformed scGen for doses >0.3 μg/kg (Figure 3E).

We used scVIDR to predict the effects of a test set of 37 drugs out of 188 treatments in the sci-Plex dose-response data[25] at 24 h for A549 cells (Figure S4A). scVIDR was trained on all data (all drugs and doses) in K562 and MCF7 cells. The model was also trained on the remaining 151 drugs in A549 cells not used in validation, as well as the vehicle data for the 37 drugs in the test set (Figure S4A). The dose-response for the 37 drugs was predicted as above by first calculating the $\hat{\delta}_{A549}$ between the control and the highest dose for a particular drug and log linearly interpolating along the $\hat{\delta}_{A549}$ in order to predict the intermediate doses. We evaluated predictions made by scVIDR at the gene, drug, and drug pathway levels. For the drug belinostat, a histone deacetylase inhibitor, scVIDR improves on predictions of DEGs such as *MALAT1* relative to scGen (Figure S4B). When predicting gene expression of the DEGs in belinostat-treated A549 cells, scVIDR also significantly outperformed scGen on all doses (Figure S4C). On predicting the DEGs of all drugs with the same mode of action as belinostat (epigenetics), scVIDR similarly outperformed scGen on all doses (Figure S4D). Finally, when looking across all 37 drugs in the test dataset, we were able to predict the expression of DEGs significantly better than scGen on average for the 3 highest doses of 100, 1,000, and 10,000 nM (Figure S4E).

Regression on the latent space infers the relationship between predicted gene expression and $\hat{\delta}_c$

Insight into model decisions can provide information regarding proper model usage and pitfalls. It would be useful to identify which genes and pathways are associated with scVIDR's prediction; however, standard VAEs do not have a linear map from the latent space to the gene expression and thus are hard to interpret. To interpret the predictions of scVIDR, we approximate the function of the decoder with linear regression (experimental procedures). We take inspiration from the use of PCA in scSeq[40] and the development of linearly decoded VAEs (LDVAEs).[41] PCA is a linear transformation that projects the data onto a lower-dimensional (latent) space while retaining as much variance as possible. This transformation is represented by a linear weight matrix, W_{pca}, with dimensions $m \times g$ where m is the number of latent variables and g is the number of genes. We can understand each principal component as a linear combination of genes. This allows us to assess the relationship between genes and a direction in latent space.

In a VAE, the mapping from the latent space to the gene space is done by the decoder that, unlike the inverse of PCA, is non-linear. In LDVAEs, however, the decoder portion of the VAE is a linear regression layer, and thus the weight matrix of this layer, W_{ldvae}, describes a linear relationship between direction in the latent space and gene prediction.[41]

However, interpretability comes at the expense of model accuracy. LDVAEs have higher reconstruction error than standard VAEs on single-cell data.[41] Similarly, using PCA and vector arithmetic to predict scSeq perturbations performed poorly compared to scGen.[35] As a result, one would like to try to interpret the latent space of a standard VAE. We present an approach to interpret the VAE's latent space using sparse regression.

We take an alternative approach to LDVAEs in which we instead approximate the non-linear function of the decoder in a standard VAE using sparse linear regression (Figure 4A). Sparse regression methods like local interpretable model-agnostic explanations (LIME) have been used to interpret complex models.[42] We specifically use sparse linear ridge regression, given that each gene has a non-zero contribution to each latent variable and that gene weights are distributed parsimoniously. This gives us a linear transformation matrix, $\widehat{W}_{vae}$, that approximates the function of the decoder.

We use this weight matrix to interrogate the relationship between predicted gene expression and $\hat{\delta}_c$. The span of $\hat{\delta}_c$ is simply a direction in scVIDR's latent space. The importance of $\hat{\delta}_c$ to each gene's predicted expression is the sum of the latent dimensional components of $\hat{\delta}_c$ multiplied by the gene's corresponding latent dimensional weight from $\widehat{W}_{vae}$. In matrix form,

$$\text{gene scores} = \hat{\delta}_c^T \widehat{W}_{vae}.$$

In practice, we found that normalizing the weight matrix by its L2 norm gives better insights when interpreting the model (experimental procedures). Gene scores represent how significant changes in latent space dimensions will impact the decoded transcriptomic response when we interpolate on the span of $\hat{\delta}_c$ on the latent space. Thus, genes with higher scores will be predicted to have bigger changes when we increase the dose of our prediction by scVIDR.

We utilize a trained scVIDR model where portal hepatocytes were left out of training and the $\hat{\delta}_{c\ =\ \text{portal hepatocytes}}$ was approximated (Figures 4B–4D). Gene scores for $\hat{\delta}_{c\ =\ \text{portal hepatocytes}}$ were calculated as described above. The genes with the top 20 highest-magnitude genes scores included well-established markers of TCDD-induced hepatotoxicity such as genes from the cytochrome P450 family (Figure 4B).[26] To see whether this relationship extended to pathways involved in TCDD-induced hepatotoxicity, we performed Enrichr analysis[38] using the 2019 WikiPathways database[43] on genes with the top 100 gene scores (Figure 4C). Among the top enriched terms, we found

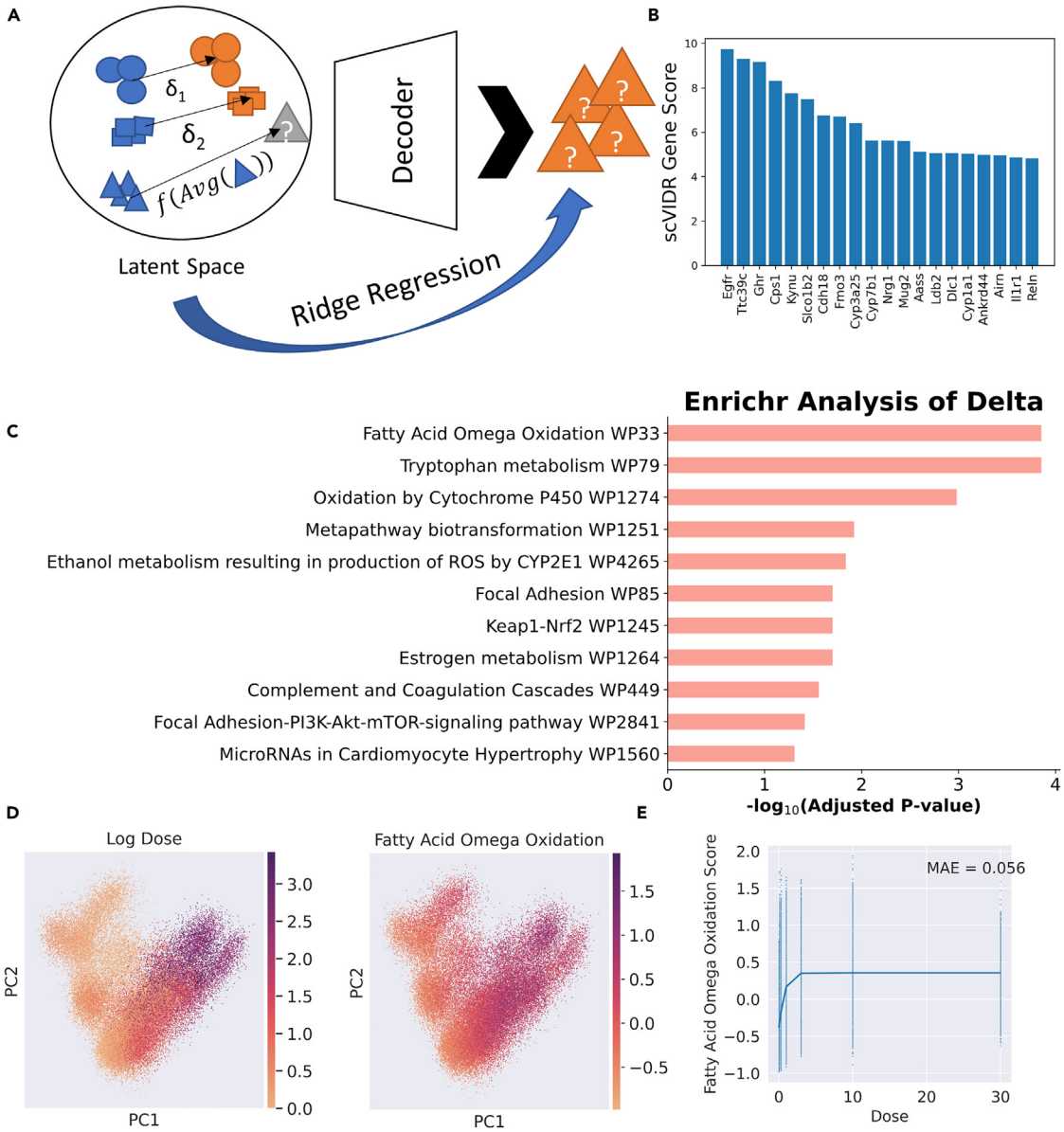


Figure 4. Interrogation of VAE using ridge regression in portal hepatocyte response prediction

(A) Schematic of calculation of latent dimension weights using ridge regression.

(B) Bar plot of top 20 genes with the highest scVIDR genes scores.

(C) Enrichr analysis of the top 100 genes with respect to the scVIDR gene scores. Bar plot of adjusted p values from statistically significant (adjusted p value < 0.05) enriched pathways from the WikiPathways 2019 Mouse Database.

(D) PCA projection of single-cell expression data colored by log dose and fatty acid oxidation pathway score.

(E) Logistic fit of median pathway score for each dose value. MAE, mean absolute error.

the hallmarks of hepatic response to TCDD in mice, such as oxidation by cytochrome P450,[44] fatty acid omega oxidation,[45] and tryptophan metabolism.[46] To derive the relationship between the actual doses and the gene pathways, the genes with the top 100 gene scores that were in "fatty acid oxidation" from WikiPathways were used in calculating enrichment scores for each cell using Scanpy.[47] A sigmoid function was fit to the median enrichment score in each dose (experimental procedures). We observed a small mean absolute error in our model and thus concluded that there was a sigmoidal dose-response relationship for the gene set generated by Enrichr (Figures 4D and 4E).

Pseudo-dose captures zonation in TCDD hepatocyte response

In single-cell analysis of developmental trajectories, it is useful to order cells with respect to a latent time course, termed "pseudo-time." This is because cells develop at different rates due to natural variations among themselves and their environment. This ordering is usually done using algorithms such as

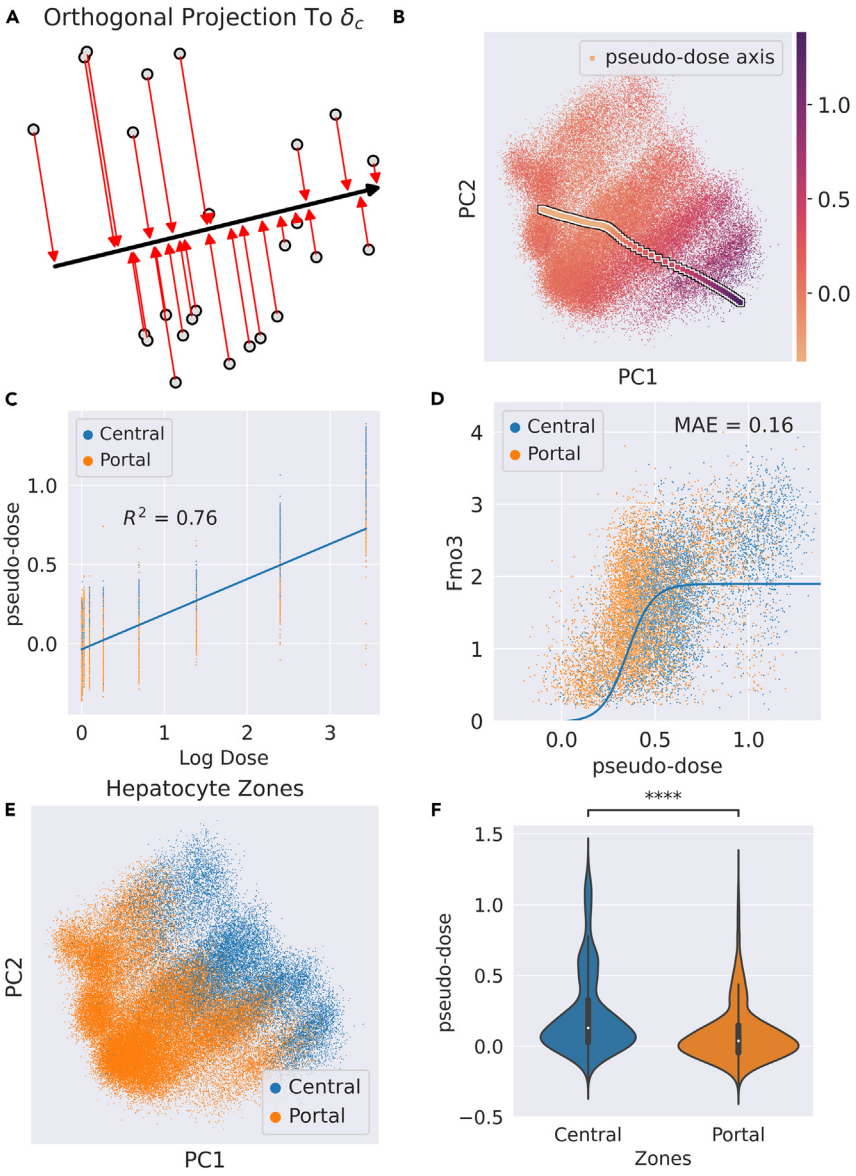


Figure 5. Pseudo-dose ordering of hepatocytes across TCDD dose-response

(A) Schematic diagram of assigning pseudo-dose values to hepatocytes by orthogonally projecting each cell in latent space to the span of the δ_c.

(B) PCA projection of hepatocytes colored by assigned pseudo-dose values. The arrow markers represent the pseudo-dose axis calculated by the δ_c.

(C) Regression plot of pseudo-dose versus log transformed real dose.

(D) Plot of pseudo-dose versus *Fmo3* expression. Associated logistic fit (solid blue line) and associated mean absolute error annotated as "MAE."

(E) PCA projection of hepatocytes colored by assigned hepatocyte zone in the liver lobule.

(F) Violin plot of the distribution of pseudo-dose values in the central and portal zones of the liver lobule. Central hepatocytes exhibit a higher pseudo-dose on average than portal hepatocytes Significance was determined by the Mann-Whitney single-sided U test. ****p < 0.0001.

Slingshot[48] and Monocle.[49] In pharmacology and toxicology, we experience a similar problem, as cells of the same type have variable sensitivities to the same toxicant. Hence, we propose to order cells in terms of a latent dose. We call this ordering of cells a "pseudo-dose."

Working off the assumption that δ_c (experimental procedures) is the axis of perturbation in latent space, we orthogonally project the latent representation of each cell to the $span(\delta_c)$ to obtain a scalar coefficient for each cell along δ_c (Figures 5A and 5B). We use this scalar coefficient as the pseudo-dose value for each cell.

To test whether these pseudo-dose values capture the latent response across cell types, we distinguished between the portal and central regions of the liver lobule. Zonation of the lobule not only defines differences in hepatocyte gene expression along the portal to the central axis but also defines their metabolic characteristics.[50] Thus, we expect that the two zones will exhibit different sensitivities to TCDD. The pseudo-dose correlated well with the actual dose administered to the hepatocytes with an R^2 = 0.76 (Figure 5C). We also found that the pseudo-dose displayed a sigmoidal relationship (experimental procedures) between the expression of DEGs such as Fmo3 (Figure 5D). Finally, we found the pseudo-dose to be statistically higher on average in the central hepatocytes versus the portal hepatocytes (Figures 5E and 5F). This is consistent with liver biology, given that central hepatocytes respond more strongly to treatment due to TCDD sequestration[51] and higher AhR expression levels in the centrilobular zone.[26]

DISCUSSION

Mapping the combinatorial space of single-cell perturbation is important to toxicology and pharmacology to facilitate the generalization of drug or toxicant effects across several domains. Computational modeling allows researchers to use current large-scale databases to predict new perturbations to scSeq data. We have demonstrated an improvement to such modeling using VAEs with regression. These improvements include highly correlated prediction of cell-type-specific effects in mouse liver, PBMCs, and A549 cells. We also modeled a latent response for mouse hepatocytes using pseudo-dose and interrogated the VAE to predict dose-dependent perturbations in portal hepatocyte pathways. We show that deep generative modeling can be used to model complex perturbations in single-cell gene expression data from several different datasets.

Model limitations

When evaluating the model in the mouse liver, scVIDR performed better on the cell types most sensitive to TCDD, e.g., hepatocytes and endothelial cells (Figures S5A, S5C, and

S5D). For cell types less sensitive to TCDD, the model often underestimated the expression of DEGs (Figure S5E). This is likely a result of a combination of factors including the similarity of the treatment to the control data (Figure S5A), the smaller control cell populations (Figure S5B), and the overall low expression of HVGs (Figure S5E). Thus, we believe that the VAE has less information to predict differential gene expression for these cell types. Our model improves on this problem with respect to scGen for most cell types in the liver (except for stellate cells and cholangiocytes at higher doses). Results from sci-Plex imply that incorporating scSeq data from livers treated with other compounds could improve these predictions, as the model would have more information on different liver responses.

In the sci-Plex dataset, prediction of certain drugs with epigenetic mode of actions produced the poorest prediction scores (Figure S6). This is because scSeq data provide no information regarding epigenetic modifications (e.g., chromatin accessibility, histone marks, and DNA-binding proteins). Integration with epigenetic data such as single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) could help to predict such responses with higher accuracy.

While scVIDR and its pseudo-dose metric work on standard dose-response scenarios, it remains untested for use with more complex cellular trajectories such as those found in development and circadian rhythms.[52] Such trajectories include branching and cycling, which involve non-linear dynamics, and may require more sophisticated models to properly capture their topology. Algorithms such as CellOT[37] can represent complex distributional shifts along latent dimensions; however, they are still only developed for single-perturbation measurements and extrapolate poorly to larger perturbations.

Future directions

When looking to the future of generative modeling in chemical-induced perturbation of gene expression, a problem domain of interest is time-dependent drug effects. Chemical exposures are not only a function of concentration but also of time.[53] Dose-time-response analysis is central to risk assessment in clinical settings.[54] Predicting the response not only as a function of amount of drug but also as a function of the time the drug is within a patient's system and the time of day at which the drug was administered would allow for more effective and safer dosing regimens.[54,55]

Developmental state can also be impacted by chemical perturbation. An example of this is the inhibition of B cell lymphopoiesis by TCDD.[56] The latent space could be useful for analyzing a simplified model of the dynamics of developmental systems and how they change with chemical perturbation. PCA for dimensionality reduction has been used in this area for successful cellular fate prediction during hematopoiesis.[57]

Conclusions

Taken together, our tool facilitates dose-response predictions for a particular drug in a specific cell type using the response of other cell types. Dose-response modeling is important in the realm of drug development and toxicity testing, as the physiological response of chemical perturbation is dose dependent. We envision the use of scVIDR in optimizing dose-response studies during drug discovery and development. scVIDR enables prediction of chemical response in a wide array of cell types and doses using only the control and the highest doses of previous experiments. As more data become available on single-cell chemical perturbations, generative modeling can yield insights into the underlying manifold of gene expression and how different classes of chemicals act on that manifold. Discovery of the properties of the manifold will allow for generalizations to be made about the physiology of tissues and understudied chemical perturbations.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this work is Sudin Bhattacharya (sbhattac@msu.edu).

Materials availability

The study did not generate new unique materials or reagents.

Data and code availability

All data used in the manuscript are publicly available and are referenced in the manuscript. The code for the software and for reproducing the figures is available at https://github.com/BhattacharyaLab/scVIDR. Long-term archive of code repository is made available via Zenodo at http://doi.org/10.5281/zenodo.8025235.[58]

Single-cell expression datasets and preprocessing

Nault et al.[23] performed all TCDD liver dose-response experiments, which were deposited in the Gene Expression Omnibus (GEO)[59] under the accession number GSE184506. Kang et al.[24] performed all IFN-β PBMC experiments, which were deposited in GEO under the accession number GSE96583. Zheng et al.[38] performed all experiments relating to study B, which were deposited in the Sequence Read Archive[60] under accession number SRP073767. Hagai et al.[39] performed all LPS6 species experiments, which were deposited in BioSciences under accession number E-MTAB-5919.[61]

The sci-Plex dataset[25] and the TCDD dose-response dataset[23] were collected and processed uniformly from raw count expression matrices. The cell expression vectors are normalized to the median total expression counts for each cell. The cell counts are then log transformed with a pseudo-count of 1. Finally, we select the top 5,000 most HVGs on which to do our analysis. The preprocessing was carried out using the *scanpy.pp* package using the *normalize_total*, *log1p*, and *highly_variable* functions.[47]

The TCDD dose-response dataset comprised of snRNA-seq of C57BL6 of flash frozen mouse livers. Mice in this dataset were administered, subchronically, a specified dose of TCDD via oral gavage every 4 days for 28 days. In our analysis, all immune cell types were left out, as immune cells are known to migrate from the lymph to the liver during TCDD administration.[22] Thus, there is a small size for the immune cell populations in the low-dose datasets versus the higher doses. PBMC data from Kang et al.,[24] study B data from Zheng et al.,[38] and species data from Hagai et al.[39] were accessed as a processed dataset from Lotfollahi et al.[35]

When training scGen and scVIDR, batch effects are accounted for with the *scvi.data* package using the *setup_anndata* function. Differential abundances of cells in different groups are accounted for by random sampling with replacement of the same number of cells for each dose and random sampling without replacement of the same number of cells for each cell type.

Implementation and training of models

All code in this manuscript is implemented in the Python programming language. The scVIDR model is built on the python package, scGen v.2.0.0,[35] which in turn is built on the python package scVI v.0.13.0.[20] Here, we modify the model to accommodate predictions of the dose-response, linear regression on the latent space, pseudo-dose calculations, and approximations of the gene importance in chemical perturbations

Hyperparameters for the model and training are the default values selected by scGen v.2.0.0. Table 1 outlines the model hyperparameters used in

Table 1. Hyperparameters for scVIDR's and scGen's variational autoencoder model

Hyperparameter	Value
Latent dimension	100
Number of layers	2
Layer width	800
Dropout rate	0.2
Kullback-Leibler weight	5e − 5

Table 2. Hyperparameters for scVIDR's and scGen's variational autoencoder training

Hyperparameter	Value
Training epochs	100
Learning rate	0.001
Learning rate decay	1e − 6
Optimizer	Adam
Optimizer epsilon	0.01
Early stopping	true
Early stopping patience	25

deploying scVIDR and scGen. Table 2 outlines the training hyperparameters when deploying scVIDR and scGen.

Our implementation of CellOT[37] and scPreGAN[36] uses default parameters from both of their respective publications.

Calculation of the δ_c for single- and multiple-dose predictions

The δ, as defined by Lotfollahi et al.,[35] is the difference between the mean latent representations of the treated (t = 1) and untreated (t = 0) conditions:

$$\delta = \bar{z}_{t=1} - \bar{z}_{t=0},$$

where $\bar{z}_t$ is the mean latent representation for treatment t in the dataset.

We can calculate a cell-type-specific $\delta_{c=A}$ for some cell type, A, by taking the difference between the mean latent representations of the treated and control groups, or

$$\delta_c = \bar{z}_{c=A,t=1} - \bar{z}_{c=A,t=0}.$$

If we want to estimate a δ_c for some type of cell type B based on $\bar{z}_{c=B,p=0}$ and where $\bar{z}_{c=B,p=1}$ is unknown, we can approximate a function based on $\bar{z}_{c=B,p=0}$, or

$$\hat{\delta}_{c=B} = f(\bar{z}_{c=B,p=0}),$$

where we approximate the above function using all other existing cell types in the dataset as input to ordinary least-squares regression as implemented by the *LinearRegression* function in the *sklearn.linear_model* package.[62]

Predictions of dose-response in the latent space in scVIDR and scGen

To predict the latent representation for a response at some dose, d, we interpolate log linearly on $\hat{\delta}_{c=B}$ such that for each latent cell in our prediction, $z_{i,c,p=d}$:

$$\hat{z}_{i,c,p=d} = z_{i,c,p=0} + \hat{\delta}_c * \frac{log(d+1)}{\log(\max(d)+1))},$$

where max (d) is the highest dose in the dataset. To calculate the dose-response values for scGen, we simply replace $\hat{\delta}_c$ with δ calculated by scGen.

Evaluating model performance

Performance of the model on the prediction task is the same as that in Lotfollahi et al.[35] We quantified performance using the R^2 value for mean gene expression for each gene across all cells. The R^2 was calculated using the *linregress* function from the *scipy.stats* package.[63] We compared the DEGs that are selected using the *rank_gene_groups* from the *Scanpy* package and taking the top 100. Models were compared on the same prediction in which we resample 80% of the cells in the cell type we are predicting 100 times. Resampling is done using the *choice* function from the *numpy.random* package.[64]

Statistical significance was determined by the one-sided Mann-Whitney U test as it is implemented by the *mannwhitneyu* function from the *scipy.stats* package. We considered p values less than 0.001 as statistically significant.

Distances were used to establish relationships between distributions and vectors. Cosine distance was calculated using the *cosine* function in the *scipy.spatial.distance* package. The Sinkhorn distance was calculated using the *SampleLoss* class in the *geomloss* package.[65]

Inferring feature-level contributions to perturbation prediction

In PCA, we perform an orthogonal linear transformation on the data such that our projected data preserve as much variance as possible. It is known that the solution to this maximization problem is to project the data onto the eigenvectors of the covariance matrix, or

$$Z_m = XW_m,$$

where X is the mean-centered scRNA-seq expression matrix, W_m is the eigenvectors corresponding to the m highest eigenvalues of the covariance matrix of X, and Z_m represents the m-dimensional projection of the data onto its principal components. We can see from this formula that Z_m is calculated as a linear combination of weights and gene expression, and thus there is a linear relationship between the genes and the principal components. We can exploit this fact and calculate a loading for each gene with each corresponding eigenvector by taking the product of the eigenvector and the square root of the corresponding eigenvalue, or

$$loading_{ij} = W_{ij} * \sqrt{\lambda_i},$$

where W_{ij} is the j^{th} value (corresponding to gene j) of the i^{th} eigenvector and λ_i is the eigenvalue for the i^{th} eigenvector. These loadings represent a normalized score of the relationship between a gene's expression and a particular principal component. These loadings are also directly proportional to the actual correlation between the gene's expression and the principal component of interest.

It can be shown that PCA and autoencoders with a single hidden layer (with a size less than the observations) and a strictly linear map are nearly equivalent.[66] We can project principal components back into expression space using the following function:

$$\hat{X} = Z_m W_m^T = XW_m W_m^T.$$

Additionally, we note that PCA is a solution to the minimization of the reconstruction error:

$$\|X - \hat{X}\|_2^2.$$

We find similarly that the loss function that we try to optimize in the autoencoder we described above is

$$\|X - XW_1 W_2^T\|_2^2,$$

where W_1 is the weights of the hidden layer and W_2 is the weights of the final layer of the autoencoder. In effect, we can see that the autoencoder described above can approximate the loadings of a PCA using W_2.

The reconstruction error for a standard VAE with the assumption that the observations are a multivariate Gaussian is

$$\frac{1}{N}\|X - Dec(Z)\|^2,$$

where N is the number of samples, $Dec(Z)$ is the function of the decoder neural network, and Z is the transformation by the encoder of the observations onto the latent space. In an LDVAE, the $Dec(Z)$ is replaced with a single

layer with linear transfer operators such that the reconstruction error is the following:

$$\frac{1}{N}\|X - ZW_{Dec}^{T}\|^{2},$$

in which W_{Dec} is the linear weights of the decoder. These weights give us an approximation of the contributions of individual genes to the dimensions of the latent space. We can interpret W_{Dec} as a loadings matrix by which we can interpret the latent dimensions of the LDVAE.

To approximate feature contributions to predicting the perturbation in scVIDR, we train a ridge regression model. We then take the decoder portion of our model and sample 100,000 points from the latent space and generate their corresponding expression vectors. This will be our training dataset for a ridge regression. We then train the ridge regression using the *Ridge* class from the *sklearn.linear_model* package. We can describe the loss of our ridge regression as

$$\|Dec(Z) - ZW^{T}\|^{2} + \lambda\|W\|^{2},$$

where Z are the sampled points from the latent space, ZW^{T} is the approximation of the predicted gene expression vectors, and W is an m × n matrix where m is the number of genes and n is the number of latent dimensions. We divide W using the $||W||_2$ to normalize for the effect of overexpressed genes. We then calculate the gene scores by taking the dot product of normalized W and δ_c, or

$$\text{gene scores} = \frac{W}{||W||_2}\cdot\delta_c.$$

We use these gene scores to order genes for *Enrichr*[67] pathway analysis with the *gseapy* package.[68] Scores for each pathway were calculated using the *score_genes* function from the *scanpy.tl* package with the genes sets derived from the *Enrichr* results.

Calculating the pseudo-dose values

We can order each cell, x_i, with respect to the variable response of x_i to the chemical by taking the latent representation, z_i, and orthogonally projecting it onto $L = span(\delta_c)$:

$$proj_L = \frac{\delta_c \cdot z_i}{\delta_c \cdot \delta_c}\delta_c = p\delta_c.$$

The scalar multiple of δ, p, is the pseudo-dose value for x_i.

Regression of sigmoid function for evaluating dose-response relationships

To establish whether a standard dose-response relationship existed between the top pathways inferred by *Enrichr* and the pseudo-dose and gene expression, a logistic function of the form

$$f(d) = \frac{L}{1+e^{-k(d-d_0)}} + b,$$

was used, where d is the dose or pseudo-dose. The parameters of the function above were fit to the output variables (median enrichment score and Fmo3 normalized expression) using the Levenberg-Marquardt algorithm implementation in the *curve_fit* function in the *scipy.optimize* package. The regression was evaluated using the mean absolute error metric implementation in the *mean_absolute_error* function in the *sklearn.metrics* package.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100817.

ACKNOWLEDGMENTS

This work was supported by the National Human Genome Research Institute R21 HG010789 to T.Z. and S.B. O.K. is supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under award number T32 ES007255. T.Z. and S.B. are partially supported by the USDA National Institute of Food and Agriculture, Michigan AgBioResearch. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported in part through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University.

AUTHOR CONTRIBUTIONS

Conceptualization, S.B. and O.K.; methodology, O.K. and D.F.; software, validation, and writing – original draft, O.K.; formal analysis, O.K., D.F., R.N., and D.M.; data curation, O.K. and R.N.; supervision and funding acquisition, S.B. and T.Z.; writing - review and editing, all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location.

Received: November 1, 2022
Revised: December 7, 2022
Accepted: July 14, 2023
Published: August 11, 2023

REFERENCES

1. Brenner, S. (2010). Sequences and consequences. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 207–212. https://doi.org/10.1098/rstb.2009.0221.
2. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell atlas. Elife *6*, e27041. https://doi.org/10.7554/eLife.27041.
3. Wilkerson, B.A., Zebroski, H.L., Finkbeiner, C.R., Chitsazan, A.D., Beach, K.E., Sen, N., Zhang, R.C., and Bermingham-Mcdonogh, O. (2021). Novel cell types and developmental lineages revealed by single-cell rna-seq analysis of the mouse crista ampullaris. Elife *10*, e60108. https://doi.org/10.7554/eLife.60108.
4. Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T.K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., et al. (2017). A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. Cell *169*, 1276–1290.e17. https://doi.org/10.1016/j.cell.2017.05.018.
5. Pellin, D., Loperfido, M., Baricordi, C., Wolock, S.L., Montepeloso, A., Weinberg, O.K., Biffi, A., Klein, A.M., and Biasco, L. (2019). A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nat. Commun. *10*, 2395. https://doi.org/10.1038/s41467-019-10291-0.
6. Rodriguez-Fraticelli, A.E., Weinreb, C., Wang, S.W., Migueles, R.P., Jankovic, M., Usart, M., Klein, A.M., Lowell, S., and Camargo, F.D. (2020). Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. Nature *583*, 585–589. https://doi.org/10.1038/s41586-020-2503-6.
7. Taylor, D.M., Aronow, B.J., Tan, K., Bernt, K., Salomonis, N., Greene, C.S., Frolova, A., Henrickson, S.E., Wells, A., Pei, L., et al. (2019). The Pediatric Cell Atlas: Defining the Growth Phase of Human Development at Single-Cell Resolution. Dev. Cell *49*, 10–29. https://doi.org/10.1016/j.devcel.2019.03.001.
8. Semrau, S., Goldmann, J.E., Soumillon, M., Mikkelsen, T.S., Jaenisch, R., and Van Oudenaarden, A. (2017). Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. Nat. Commun. *8*, 1096. https://doi.org/10.1038/s41467-017-01076-4.
9. van Galen, P., Hovestadt, V., Wadsworth Ii, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J., et al. (2019). Single-Cell RNA-Seq Reveals AML Hierarchies

Relevant to Disease Progression and Immunity. Cell *176*, 1265–1281.e24. https://doi.org/10.1016/j.cell.2019.01.031.

10. Peng, J., Sun, B.F., Chen, C.Y., Zhou, J.Y., Chen, Y.S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G.S., et al. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. *29*, 725–738. https://doi.org/10.1038/s41422-019-0195-y.
11. Brivanlou, A.H., and Darnell, J.E. (2002). Signal Transduction and the Control of Gene Expression. Science *295*, 813–818. https://doi.org/10.1126/science.1066355.
12. Blumenthal, D.K. (2017). Pharmacodynamics: Molecular Mechanisms of Drug Action. In Goodman & Gilman's: The Pharmacological Basis of Therapeutics, 13e, L.L. Brunton, R. Hilal-Dandan, and B.C. Knollmann, eds. (McGraw-Hill Education).
13. Yao, J., Pilko, A., and Wollman, R. (2016). Distinct cellular states determine calcium signaling response. Mol. Syst. Biol. *12*, 894. https://doi.org/10.15252/MSB.20167137.
14. Kramer, B.A., and Pelkmans, L. (2019). Cellular state determines the multimodal signaling response of single cells. Preprint at bioRxiv. https://doi.org/10.1101/2019.12.18.880930.
15. Zhang, Q., Caudle, W.M., Pi, J., Bhattacharya, S., Andersen, M.E., Kaminski, N.E., and Conolly, R.B. (2019). Embracing systems toxicology at single-cell resolution. Curr. Opin. Toxicol. *16*, 49–57. https://doi.org/10.1016/j.cotox.2019.04.003.
16. Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Ji, Y., Ibarra, I.-C.L., Wolf, F.A., Yakubova, N., Theis, F.J., and Lopez-Paz, D. (2021). Learning interpretable cellular responses to complex perturbations in high-throughput screens. Preprint at bioRxiv. https://doi.org/10.1101/2021.04.14.439903.
17. Peidli, S., Green, T.D., Shen, C., Gross, T., Min, J., Garda, S., Yuan, B., Schumacher, L.J., Taylor-King, J., Marks, D., et al. (2023). scPerturb: Harmonized Single-Cell Perturbation Data. Preprint at bioRxiv. https://doi.org/10.1101/2022.08.20.504663.
18. McFarland, J.M., Paolella, B.R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Colgan, W.N., Jones, A., Chambers, E., et al. (2020). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nat. Commun. *11*, 4296. https://doi.org/10.1038/s41467-020-17440-w.
19. Kingma, D.P., and Welling, M. (2014). Auto-encoding variational bayes. In 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (International Conference on Learning Representations (ICLR).
20. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.
21. Qiu, Y.L., Zheng, H., and Gevaert, O. (2020). Genomic data imputation with variational auto-encoders. GigaScience *9*. giaa082–12. https://doi.org/10.1093/gigascience/giaa082.
22. Nault, R., Fader, K.A., Bhattacharya, S., and Zacharewski, T.R. (2021). Single-Nuclei RNA Sequencing Assessment of the Hepatic Effects of 2,3,7,8-Tetrachlorodibenzo-p-dioxin. CMGH *11*, 147–159. https://doi.org/10.1016/j.jcmgh.2020.07.012.
23. Nault, R., Saha, S., Bhattacharya, S., Dodson, J., Sinha, S., Maiti, T., and Zacharewski, T. (2022). Benchmarking of a Bayesian single cell RNAseq differential gene expression test for dose–response study designs. Nucleic Acids Res. *50*, e48. https://doi.org/10.1093/nar/gkac019.
24. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol. *36*, 89–94. https://doi.org/10.1038/nbt.4042.
25. Srivatsan, S.R., McFaline-Figueroa, J.L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H.A., Jackson, D.L., Daza, R.M., Christiansen, L., et al. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. Science *367*, 45–51. https://doi.org/10.1126/science.aax6234.
26. Lindros, K.O., Oinonen, T., Johansson, I., and Ingelman-Sundberg, M. (1997). Selective Centrilobular Expression of the Aryl Hydrocarbon Receptor in Rat Liver. J. Pharmacol. Exp. Therapeut. *280*, 506–511.
27. Yang, Y., Filipovic, D., and Bhattacharya, S. (2022). A Negative Feedback Loop and Transcription Factor Cooperation Regulate Zonal Gene Induction by 2, 3, 7, 8-Tetrachlorodibenzo-p-Dioxin in the Mouse Liver. Hepatol. Commun. *6*, 750–764. https://doi.org/10.1002/hep4.1848.
28. Fefferman, C., Mitter, S., and Narayanan, H. (2016). Testing the manifold hypothesis. J. Am. Math. Soc. *29*, 983–1049. https://doi.org/10.1090/jams/852.
29. Davidson, E.H. (2006). The "Regulatory Genome" for Animal Development. In The Regulatory Genome (Elsevier), pp. 1–29. https://doi.org/10.1016/b978-012088563-3.50019-5.
30. Sun, S., Zhu, J., Ma, Y., and Zhou, X. (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. *20*, 269. https://doi.org/10.1186/s13059-019-1898-6.
31. Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. Nat. Commun. *11*, 1201. https://doi.org/10.1038/s41467-020-14766-3.
32. Ding, J., Condon, A., and Shah, S.P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat. Commun. *9*, 2002. https://doi.org/10.1038/s41467-018-04368-5.
33. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 390. https://doi.org/10.1038/s41467-018-07931-2.
34. Grønbech, C.H., Vording, M.F., Timshel, P.N., Sønderby, C.K., Pers, T.H., and Winther, O. (2020). scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics *36*, 4415–4422. https://doi.org/10.1093/bioinformatics/btaa293.
35. Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721. https://doi.org/10.1038/s41592-019-0494-8.
36. Wei, X., Dong, J., and Wang, F. (2022). scPreGAN, a deep generative model for predicting the response of single-cell expression to perturbation. Bioinformatics *38*, 3377–3384. https://doi.org/10.1093/bioinformatics/btac357.
37. Bunne, C., Stark, S.G., Gut, G., del Castillo, J.S., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. (2021). Learning Single-Cell Perturbation Responses using Neural Optimal Transport. Preprint at bioRxiv. https://doi.org/10.1101/2021.12.15.472775.
38. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049–14112. https://doi.org/10.1038/ncomms14049.
39. Hagai, T., Chen, X., Miragaia, R.J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.-E., Proserpio, V., Donati, G., et al. (2018). Gene expression variability across cells and species shapes innate immunity. Nature *563*, 197–202.
40. Rostom, R., Svensson, V., Teichmann, S.A., and Kar, G. (2017). Computational approaches for interpreting scRNA-seq data. FEBS Lett. *591*, 2213–2225. https://doi.org/10.1002/1873-3468.12684.
41. Svensson, V., Gayoso, A., Yosef, N., and Pachter, L. (2020). Interpretable factor models of single-cell RNA-seq via variational autoencoders. Bioinformatics *36*, 3418–3421. https://doi.org/10.1093/bioinformatics/btaa169.
42. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.
43. Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., A Miller, R., Digles, D., Lopes, E.N., Ehrhart, F., et al. (2021). WikiPathways: Connecting communities. Nucleic Acids Res. *49*, D613–D621. https://doi.org/10.1093/nar/gkaa1024.

44. Henry, E.C., Welle, S.L., and Gasiewicz, T.A. (2010). TCDD and a Putative Endogenous AhR Ligand, ITE, Elicit the Same Immediate Changes in Gene Expression in Mouse Lung Fibroblasts. Toxicol. Sci. *114*, 90–100. https://doi.org/10.1093/toxsci/kfp285.

45. Cholico, G.N., Fling, R.R., Zacharewski, N.A., Fader, K.A., Nault, R., and Zacharewski, T.R. (2021). Thioesterase induction by 2,3,7,8-tetrachlorodibenzo-p-dioxin results in a futile cycle that inhibits hepatic β-oxidation. Sci. Rep. *11*, 15689. https://doi.org/10.1038/s41598-021-95214-0.

46. Friedrich, M., Sankowski, R., Bunse, L., Kilian, M., Green, E., Ramallo Guevara, C., Pusch, S., Poschet, G., Sanghvi, K., Hahn, M., et al. (2021). Tryptophan metabolism drives dynamic immunosuppressive myeloid states in IDH-mutant gliomas. Nat. Can. (Que.) *2*, 723–740. https://doi.org/10.1038/s43018-021-00201-z.

47. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.

48. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom. *19*, 477. https://doi.org/10.1186/s12864-018-4772-0.

49. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982. https://doi.org/10.1038/nmeth.4402.

50. Cunningham, R.P., and Porat-Shliom, N. (2021). Liver Zonation – Revisiting Old Questions With New Technologies. Front. Physiol. *12*, 732929. https://doi.org/10.3389/FPHYS.2021.732929.

51. Santostefano, M.J., Richardson, V.M., Walker, N.J., Blanton, J., Lindros, K.O., Lucier, G.W., Alcasey, S.K., and Birnbaum, L.S. (1999). Dose-dependent localization of TCDD in isolated centrilobular and periportal hepatocytes. Toxicol. Sci. *52*, 9–19. https://doi.org/10.1093/toxsci/52.1.9.

52. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. https://doi.org/10.1038/s41587-019-0071-9.

53. Lioy, P.J. (1990). Assessing total human exposure to contaminants: A multidisciplinary approach. Environ. Sci. Technol. *24*, 938–945. https://doi.org/10.1021/es00077a001.

54. Gabrielsson, J., Andersson, R., Jirstrand, M., and Hjorth, S. (2019). Dose-Response-Time Data Analysis: An Underexploited Trinity. Pharmacol. Rev. *71*, 89–122. https://doi.org/10.1124/pr.118.015750.

55. Dobrek, L. (2021). Chronopharmacology in Therapeutic Drug Monitoring—Dependencies between the Rhythmics of Pharmacokinetic Processes and Drug Concentration in Blood. Pharmaceutics *13*. https://doi.org/10.3390/pharmaceutics13111915.

56. Li, J., Bhattacharya, S., Zhou, J., Phadnis-Moghe, A.S., Crawford, R.B., and Kaminski, N.E. (2017). Aryl Hydrocarbon Receptor Activation Suppresses EBF1 and PAX5 and Impairs Human B Lymphopoiesis. J. Immunol. *199*, 3504–3515. https://doi.org/10.4049/jimmunol.1700289.

57. Yeo, G.H.T., Saksena, S.D., and Gifford, D.K. (2021). Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. Nat. Commun. *12*, 3222. https://doi.org/10.1038/s41467-021-23518-w.

58. Kana, O.Z.; BhattacharyaLab (2023). BhattacharyaLab/scVIDR: Gamma. https://doi.org/10.5281/ZENODO.8025235.

59. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*, D991–D995. https://doi.org/10.1093/nar/gks1193.

60. Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J.R., and O'Sullivan, C. (2022). The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. *50*, D387–D390. https://doi.org/10.1093/nar/gkab1053.

61. Hagai, T. (2018). RNA-seq of of dermal fibroblasts.

62. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

63. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

64. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

65. Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S., Trouvé, A., and Peyré, G. (2018). Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. Preprint at arXiv. https://doi.org/10.48550/arxiv.1810.08278.

66. Plaut, E. (2018). From principal subspaces to principal components with linear autoencoders. Preprint at arXiv. https://doi.org/10.48550/arXiv.1804.10253.

67. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinf. *14*, 128. https://doi.org/10.1186/1471-2105-14-128.

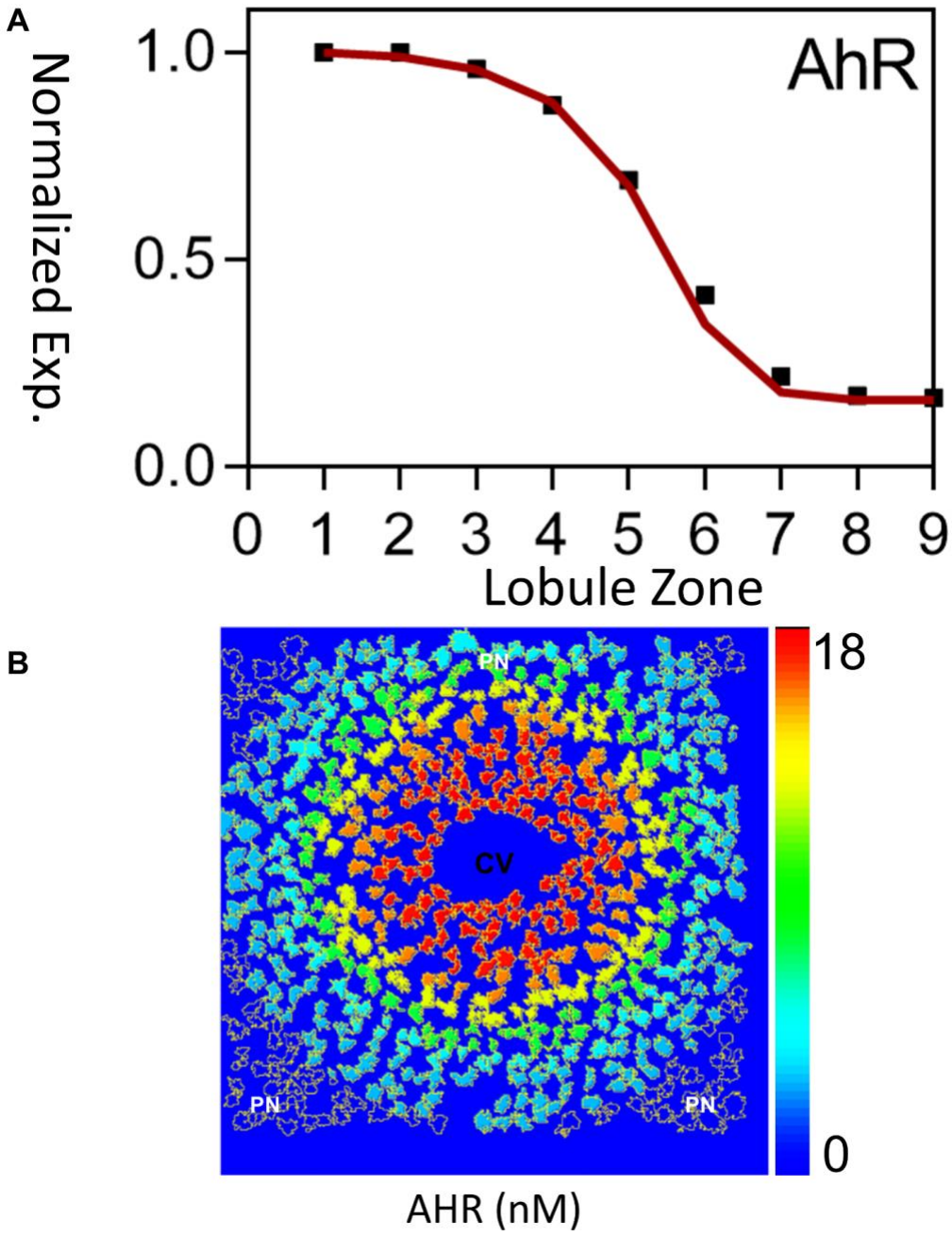
68. Fang, Z., Liu, X., and Peltz, G. (2023). GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics *39*, btac757. https://doi.org/10.1093/bioinformatics/btac757.

Patterns, Volume *4*

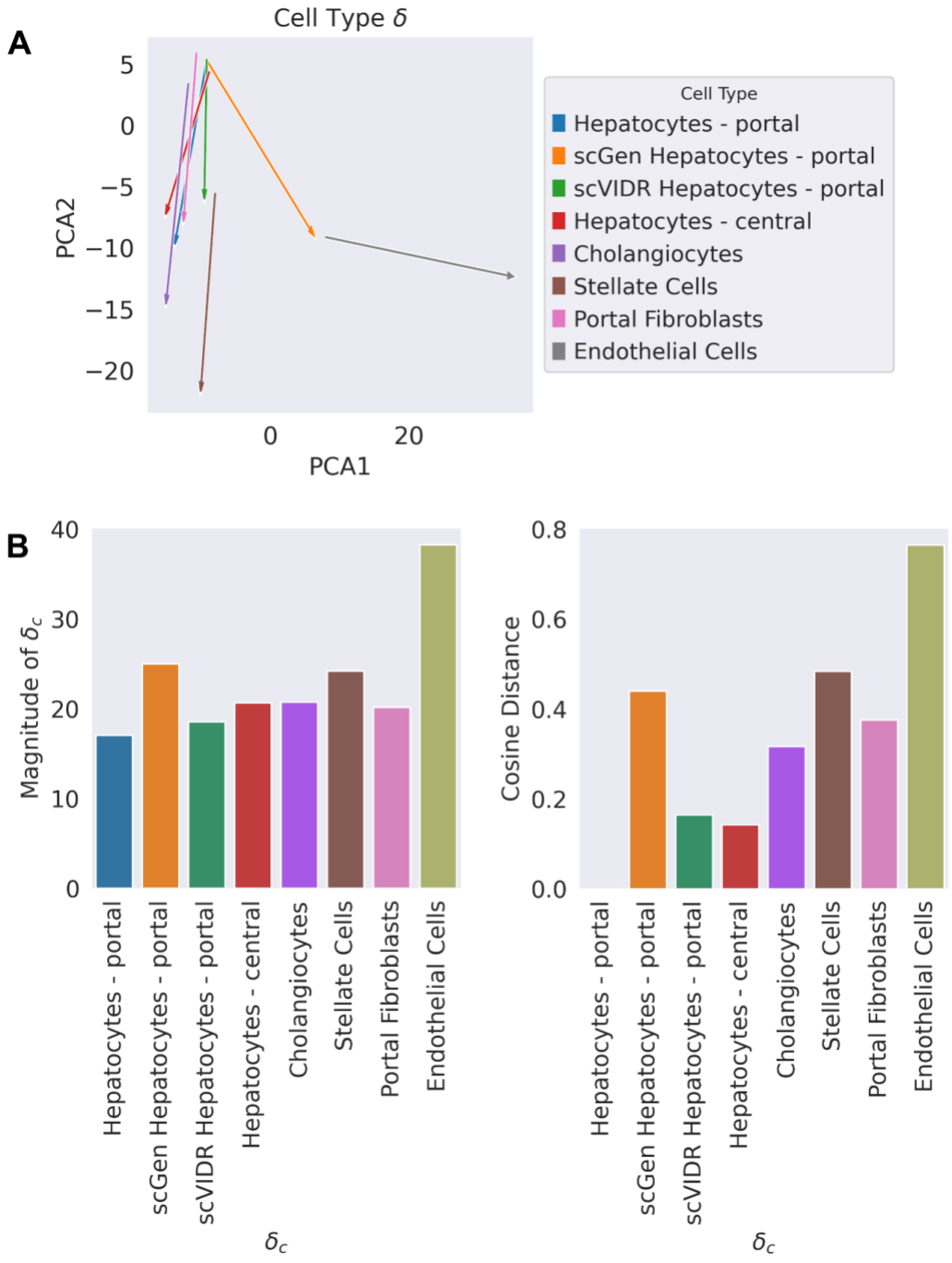
Supplemental information

Generative modeling of single-cell gene expression for dose-dependent chemical perturbations

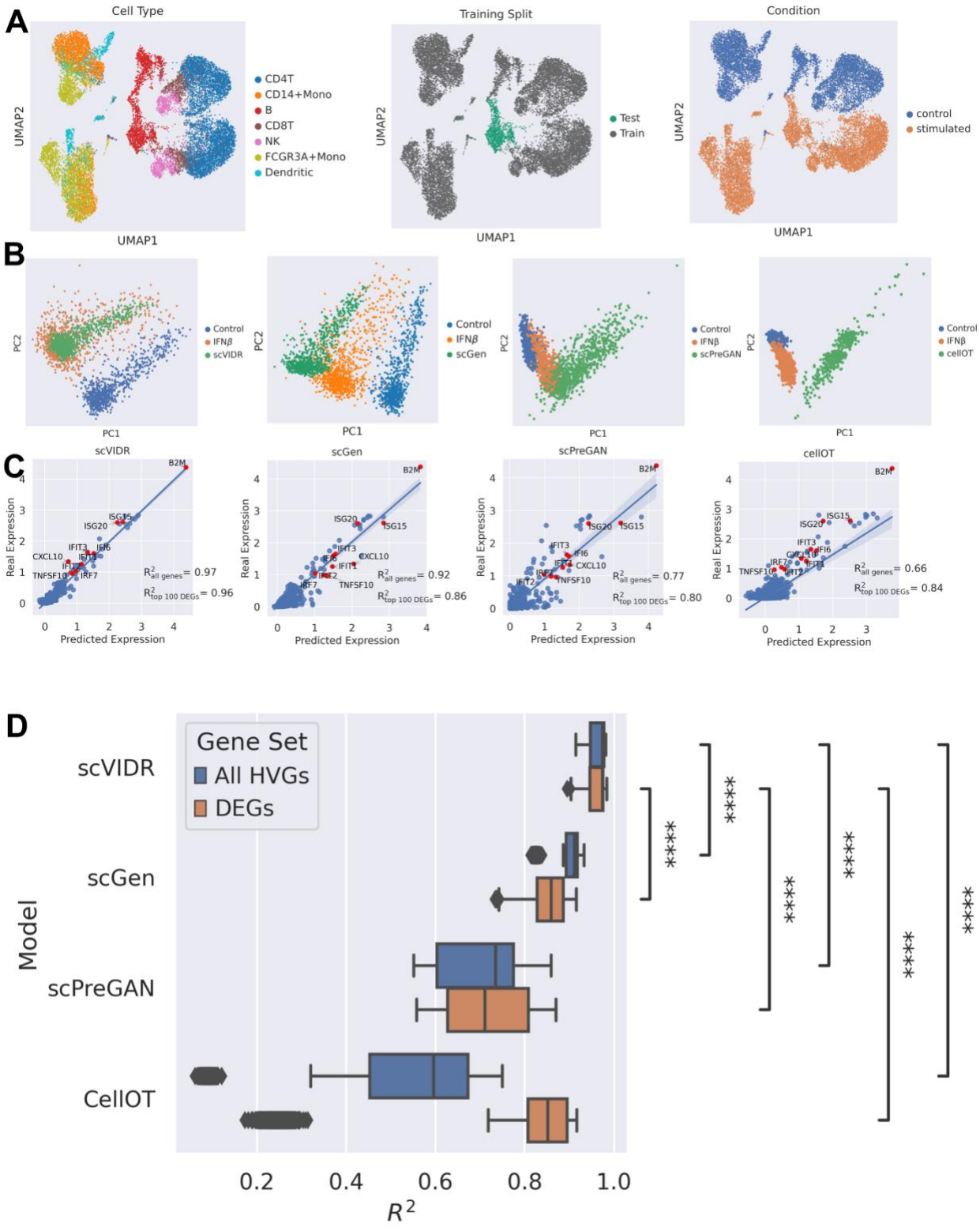
Omar Kana, Rance Nault, David Filipovic, Daniel Marri, Tim Zacharewski, and Sudin Bhattacharya



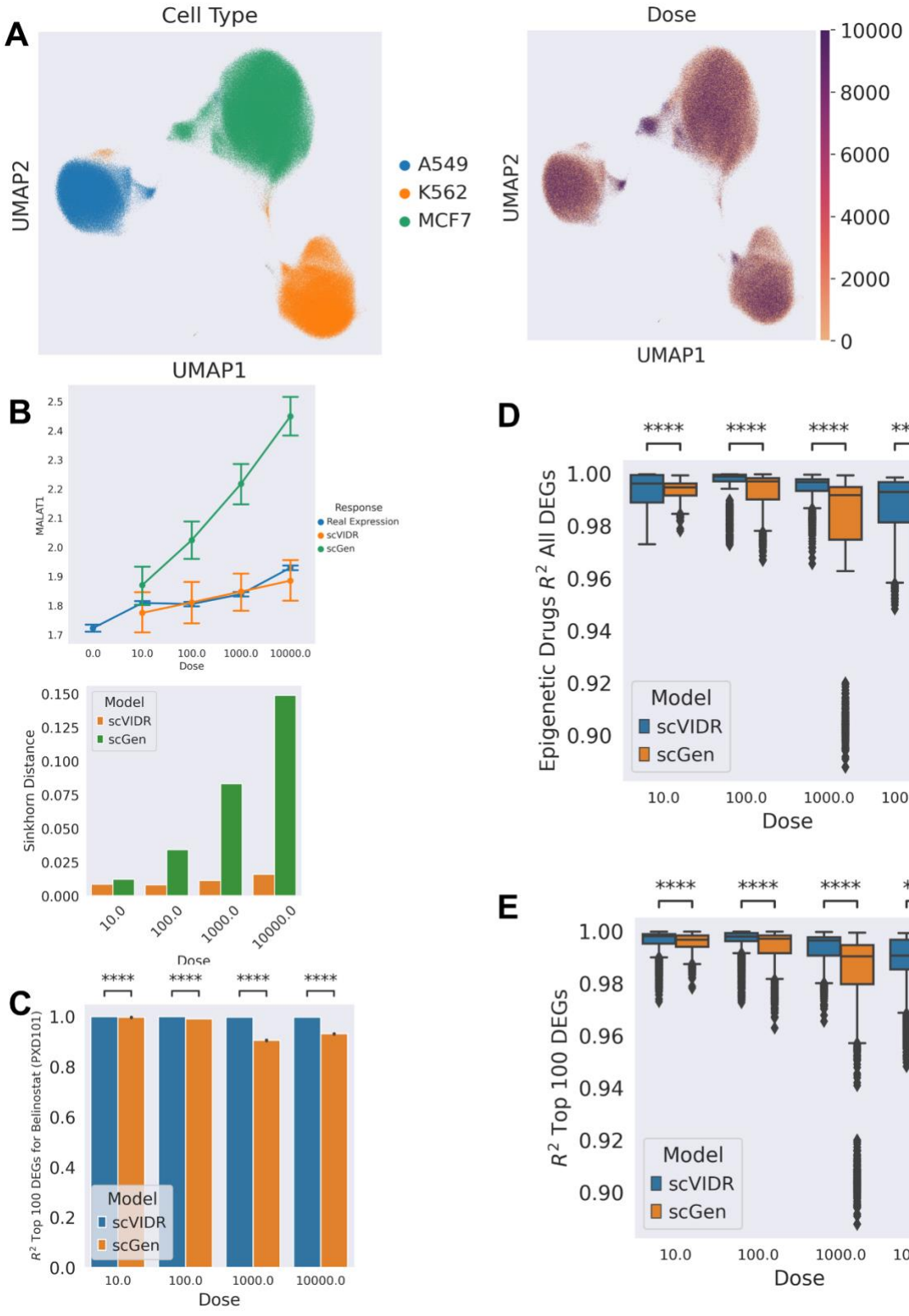
Supplementary Figure 1. Expression of AhR across liver lobule A) A zonal expression profile of normalized expression as described by Yang et al.[1] and Halpern et al.[2] Zone 0 represents the level of AhR expression in hepatocytes closest to the central vein. Zone 9 represents the level of AhR expression closest to the portal vein. **B)** A single cell resolution image of the liver lobule generated by Halpern et al. with expression levels represented by color from Yang et al. The central vein is denoted by "CV" (black) with the portal triad denoted by "PN" (white).



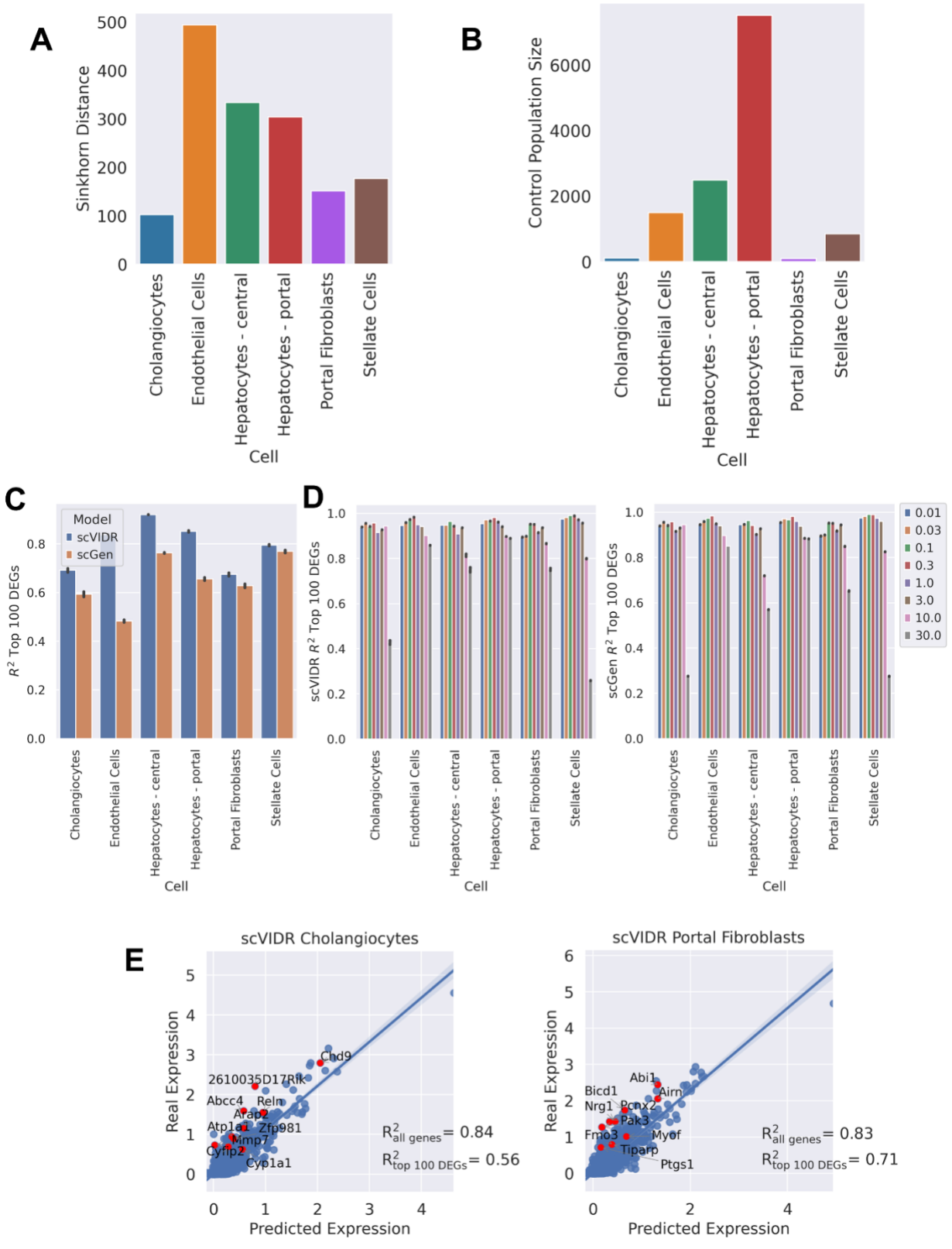
Supplementary Figure 2. δ_{scGen} deviates more from $\delta_{Hepatocytes-portal}$ than δ_{scVIDR} . A) A PCA visualization of the calculated δ_cs for a VAE trained without portal hepatocytes. "scGen Hepatocytes – portal" refers to the prediction by scGen (δ_{scGen})**,** and "scVIDR Hepatocytes – portal" refers to the prediction by scVIDR (δ_{scVIDR}) . **B)** Bar plots of the magnitude of the $\delta_c s$, and the cosine distance from the $\delta_{Hepatocytes-portal}$ for each δ_c. A cosine distance of 0 represents a δ_c in the same direction as $\delta_{Hepatocytes-portal}$, of 1 represents a δ_c orthogonal to $\delta_{Hepatocytes-portal}$ and of 2 represent a δ_c in the opposite direction as $\delta_{Hepatocytes-portal}$.


A
Cell Type
UMAP2
UMAP1
CD4T
CD14+Mono
B
CD8T
NK
FCGR3A+Mono
Dendritic
Training Split
Test
Train
Condition
control
stimulated
B
PC2
PC1
Control
IFNβ
scVIDR
scGen
scPreGAN
cellOT
C
scVIDR
scGen
scPreGAN
cellOT
Real Expression
Predicted Expression
B2M
ISG15
ISG20
IFIT3
IFI6
CXCL10
IFIT1
IRF7
IRF2
IFIT2
TNFSF10
$R^2_{\text{all genes}}$= 0.97
$R^2_{\text{top 100 DEGs}}$= 0.96
$R^2_{\text{all genes}}$= 0.92
$R^2_{\text{top 100 DEGs}}$= 0.86
$R^2_{\text{all genes}}$= 0.77
$R^2_{\text{top 100 DEGs}}$= 0.80
$R^2_{\text{all genes}}$= 0.66
$R^2_{\text{top 100 DEGs}}$= 0.84
D
Gene Set
All HVGs
DEGs
Model
scVIDR
scGen
scPreGAN
CellOT
R^2
****

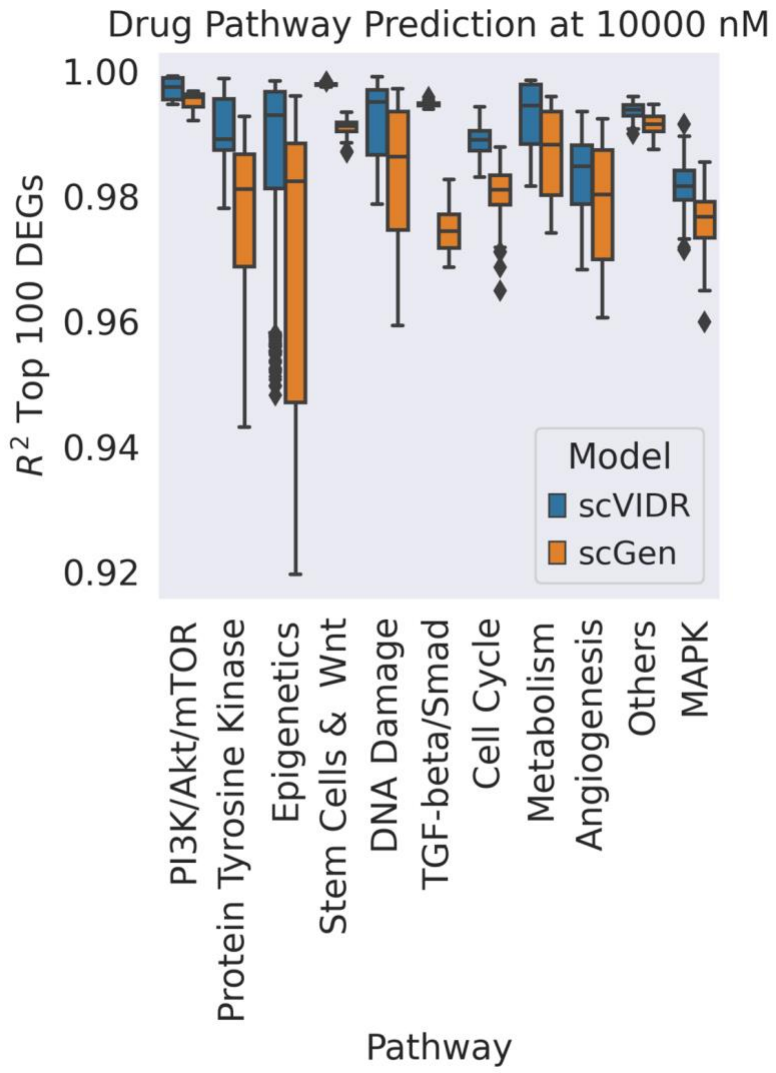
Supplementary Figure 3. Prediction of in vitro response of B-cells to IFNβ. A) UMAP of latent space of treated and untreated single-cell expression. UMAP plots are colored by cell type, training split, and condition, respectively**. B)** PCA plot of scGen, scVIDR, scPreGAN, and CellOT predictions of B-cell expression after IFNβ treatment. **C)** scGen, scVIDR, scPreGAN, and CellOT prediction versus experimental expression data regression plot. Each point represents the mean expression for a particular gene. Red points represent the top ten differentially expressed genes. Shaded region around regression line represents the 95% confidence interval. **D)** Boxplot of R^2 scores across all tissues in the PBMC treated dataset. Prediction of all highly variable genes (blue), and top 100 differentially expressed genes (orange).


A
Cell Type
UMAP2
UMAP1
A549
K562
MCF7
Dose
10000
8000
6000
4000
2000
0
B
MALAT1
Response
Real Expression
scVIDR
scGen
Dose
Sinkhorn Distance
Model
scVIDR
scGen
C
R^2 Top 100 DEGs for Belinostat (PXD101)
****
D
Epigenetic Drugs R^2 All DEGs
E
R^2 Top 100 DEGs

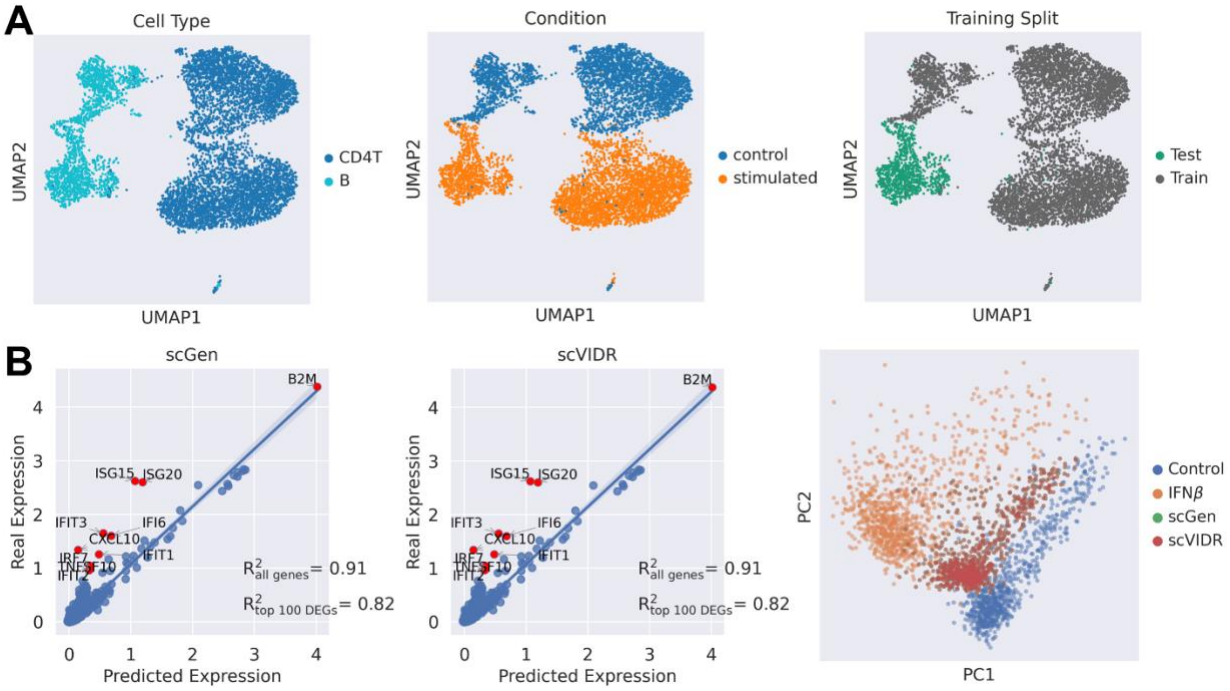
Supplementary Figure 4. Prediction of in vitro dose-response of A549 cells to different drug treatments. A) UMAP of the latent space of single-cell expression colored by cell type and dose (nM) respectively. **B)** Prediction of the dose-response of MALAT1 in response to Belinostat treatment of A549 cells. The differences between the predicted and true distribution and of MALAT1 at each dose are measured via the Sinkhorn distance. **C)** Bar plot of prediction performance of the dose-response of Belinostat administered to A549 cells on the top 100 differentially expressed genes **D)** Boxplot of prediction performance of the top 100 differentially expressed genes for the A549 dose-response in all test dataset epigenetic pathway drugs. **E)** Boxplot of prediction performance of the top 100 differentially expressed for the A549 dose-response in all 37 test dataset drugs.


A
Sinkhorn Distance
0
100
200
300
400
500
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Portal Fibroblasts
Stellate Cells
Cell
B
Control Population Size
0
2000
4000
6000
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Portal Fibroblasts
Stellate Cells
Cell
C
Model
scVIDR
scGen
R^2 Top 100 DEGs
0.0
0.2
0.4
0.6
0.8
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Portal Fibroblasts
Stellate Cells
Cell
D
scVIDR R^2 Top 100 DEGs
scGen R^2 Top 100 DEGs
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.03
0.1
0.3
1.0
3.0
10.0
30.0
Cholangiocytes
Endothelial Cells
Hepatocytes - central
Hepatocytes - portal
Portal Fibroblasts
Stellate Cells
Cell
E
scVIDR Cholangiocytes
Real Expression
Predicted Expression
Chd9
2610035D17Rik
Abcc4
Reln
Arap2
Atp1a1
Zfp981
Mmp7
Cyfip2
Cyp1a1
$R^2_{\text{all genes}}$= 0.84
$R^2_{\text{top 100 DEGs}}$= 0.56
scVIDR Portal Fibroblasts
Real Expression
Predicted Expression
Abi1
Bicd1
Airn
Nrg1
Pcnx2
Pak3
Fmo3
Myof
Tiparp
Ptgs1
$R^2_{\text{all genes}}$= 0.83
$R^2_{\text{top 100 DEGs}}$= 0.71

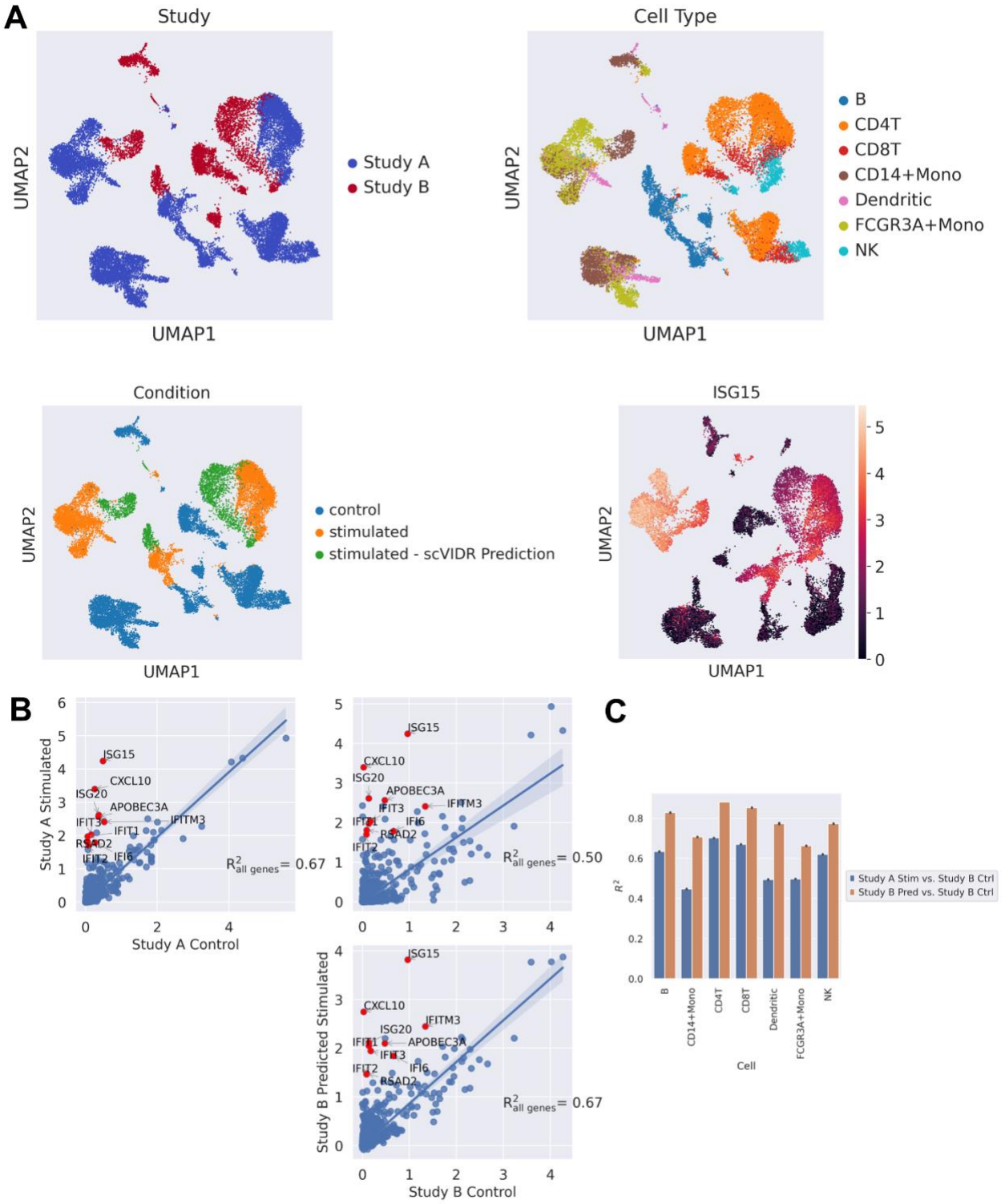
Supplementary Figure 5. Impact of latent perturbation magnitude and control population size on overall model performance. A) Sinkhorn distance between the latent distributions of the control and 30 μg/kg doses of TCDD of each cell type on the latent space. **B)** Bar plot of the control group cell population size for each cell type. **C)** Bar plot of mean gene R^2 for each individual cell type when predicting only the 30 μg/kg dose of TCDD. **D)** Bar plot of mean R^2 for each individual cell type when predicting across the entire TCDD dose-response experiment**. E)** scVIDR prediction versus real expression regression plot of cholangiocytes and stellate cell from mice administered with a 30 μg/kg dose of TCDD. Each point represents the mean expression of a gene. The top 10 differentially expressed genes are represented with red points.



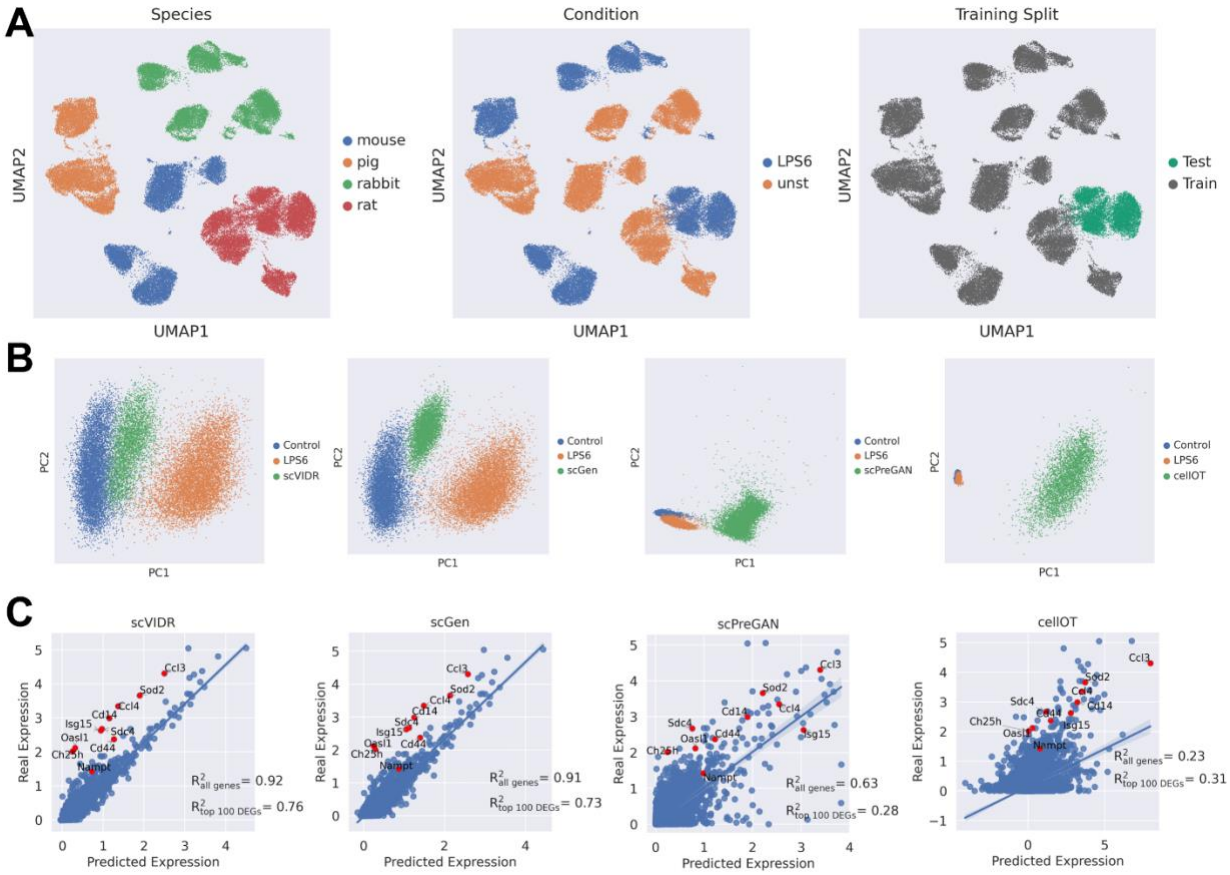
Supplementary Figure 6. Overall drug pathway performances at the highest administered dose in sci-plex dataset. A) A boxplot of the mean gene R^2 across all drug pathways in the test dataset at a dose of 10,000 nM.



Supplementary Figure 7. scVIDR is equivalent to scGen when training on a single cell type. A) A UMAP of latent space of single-cell expression of two cell types from Kang et al[3]: CD4T and B cells. They are colored by cell type, condition, and train test split. **B)** Validaton of prediction of B-cell perturbation when VAE is trained solely on CD4-T cells. A regression plot is shown for both scVIDR and scGen performance. Each point represents the mean expression of a particular gene. Red points represent the top ten differentially expressed genes. Shaded region around regression line represents the 95% confidence interval.


A
Study
UMAP2
UMAP1
Study A
Study B
Cell Type
B
CD4T
CD8T
CD14+Mono
Dendritic
FCGR3A+Mono
NK
Condition
control
stimulated
stimulated - scVIDR Prediction
ISG15
B
Study A Stimulated
Study A Control
Study B Predicted Stimulated
Study B Control
ISG15
CXCL10
ISG20
APOBEC3A
IFIT3
IFITM3
IFIT1
RSAD2
IFIT2
IFI6
$R^2_{\text{all genes}}$ = 0.67
$R^2_{\text{all genes}}$ = 0.50
$R^2_{\text{all genes}}$ = 0.67
C
R^2
Cell
B
CD14+Mono
CD4T
CD8T
Dendritic
FCGR3A+Mono
NK
Study A Stim vs. Study B Ctrl
Study B Pred vs. Study B Ctrl

Supplementary Figure 8. scVIDR exhibits similar capabilities to scGEN when doing cross-study predictions. A) A UMAP of the latent space of single-cell expression from two studies: Kang et al[3] (Study A) and Zheng et al[4] (Study B). Study B perturbation by IFN-β was predicted by scVIDR. The cells are colored by study, cell type, condition/prediction, and ISG15 expression. **B)** A regression plot comparing Study A with Study B in terms of FGRC+Mono cells. Each point represents the mean expression of a particular gene. Red points represent the top ten differentially expressed genes. Shaded region around line represents the 95% confidence interval. **C)** A barplot representing the correlation between Study A cells stimulated by IFN-β with Study B control and the correlation between scVIDR predicted Study B cells stimulated by IFN-β and Study B control.



Supplementary Figure 9. scVIDR predicts the effects of LPS6 on rat cells from mouse, rabbit, and pig cells better than other state-of-the-art algorithms. A) UMAP of latent space of treated and untreated single-cell expression. UMAP plots are colored by species, training split, and condition, respectively**. B)** PCA plot of scGen, scVIDR, scPreGAN, and CellOT predictions of rat after LPS6 treatment. **C)** scGen, scVIDR, scPreGAN, and CellOT prediction versus experimental expression data regression plot. Each point represents the mean expression for a particular gene. Red points represent the top ten differentially expressed genes. Shaded region around regression line represents the 95% confidence interval.

References

1. Yang, Y., Filipovic, D., and Bhattacharya, S. (2021). A Negative Feedback Loop and Transcription Factor Cooperation Regulate Zonal Gene Induction by 2, 3, 7, 8-Tetrachlorodibenzo-p-Dioxin in the Mouse Liver. Hepatol. Commun. 10.1002/hep4.1848.

2. Halpern, K.B., Shenhav, R., Matcovitch-Natan, O., Tóth, B., Lemze, D., Golan, M., Massasa, E.E., Baydatch, S., Landen, S., Moor, A.E., et al. (2017). Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nat. 2017 5427641 *542*, 352–356. 10.1038/nature21065.

3. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol. *36*, 89–94. 10.1038/nbt.4042.

4. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 1–12. 10.1038/ncomms14049.