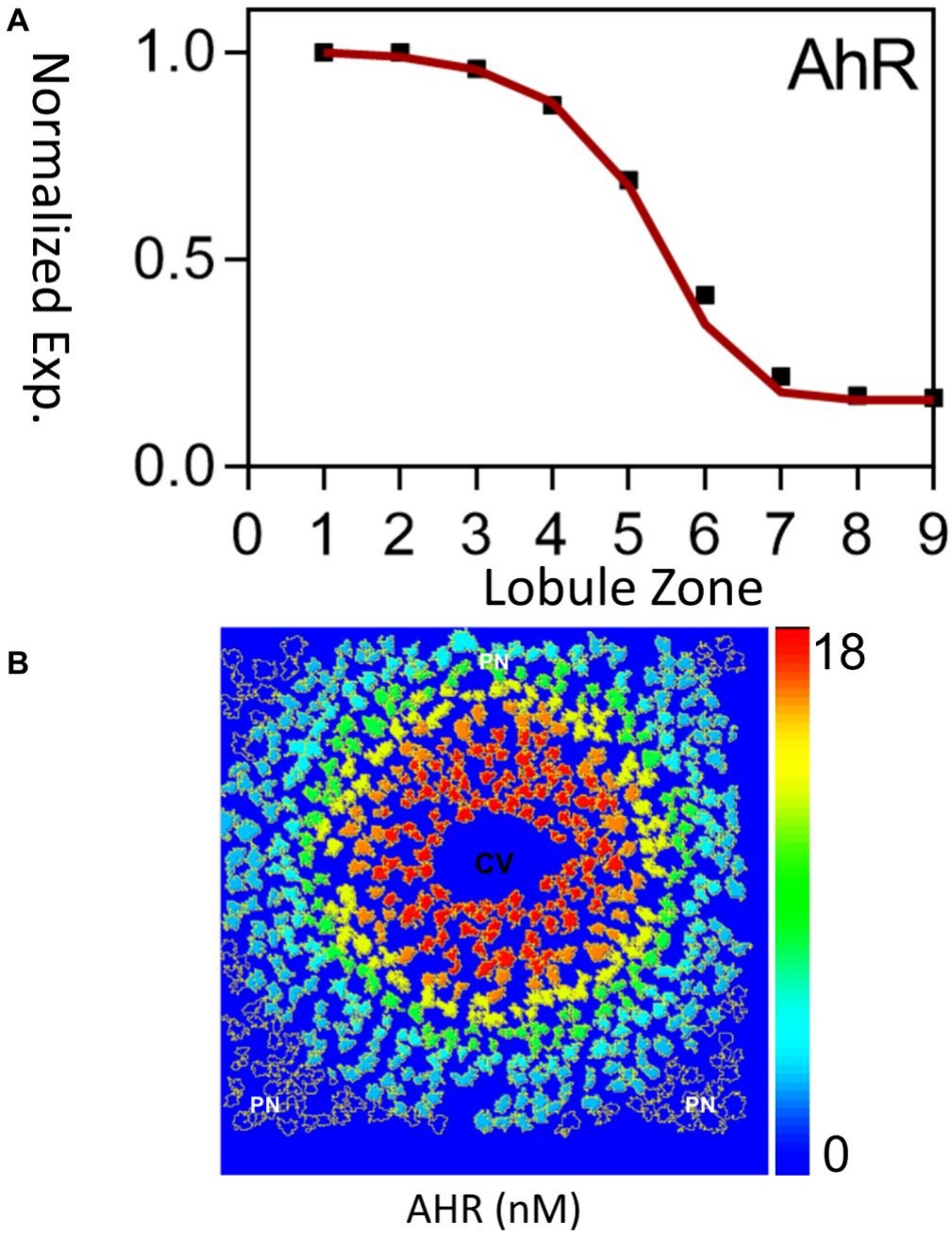
Patterns, Volume *4*

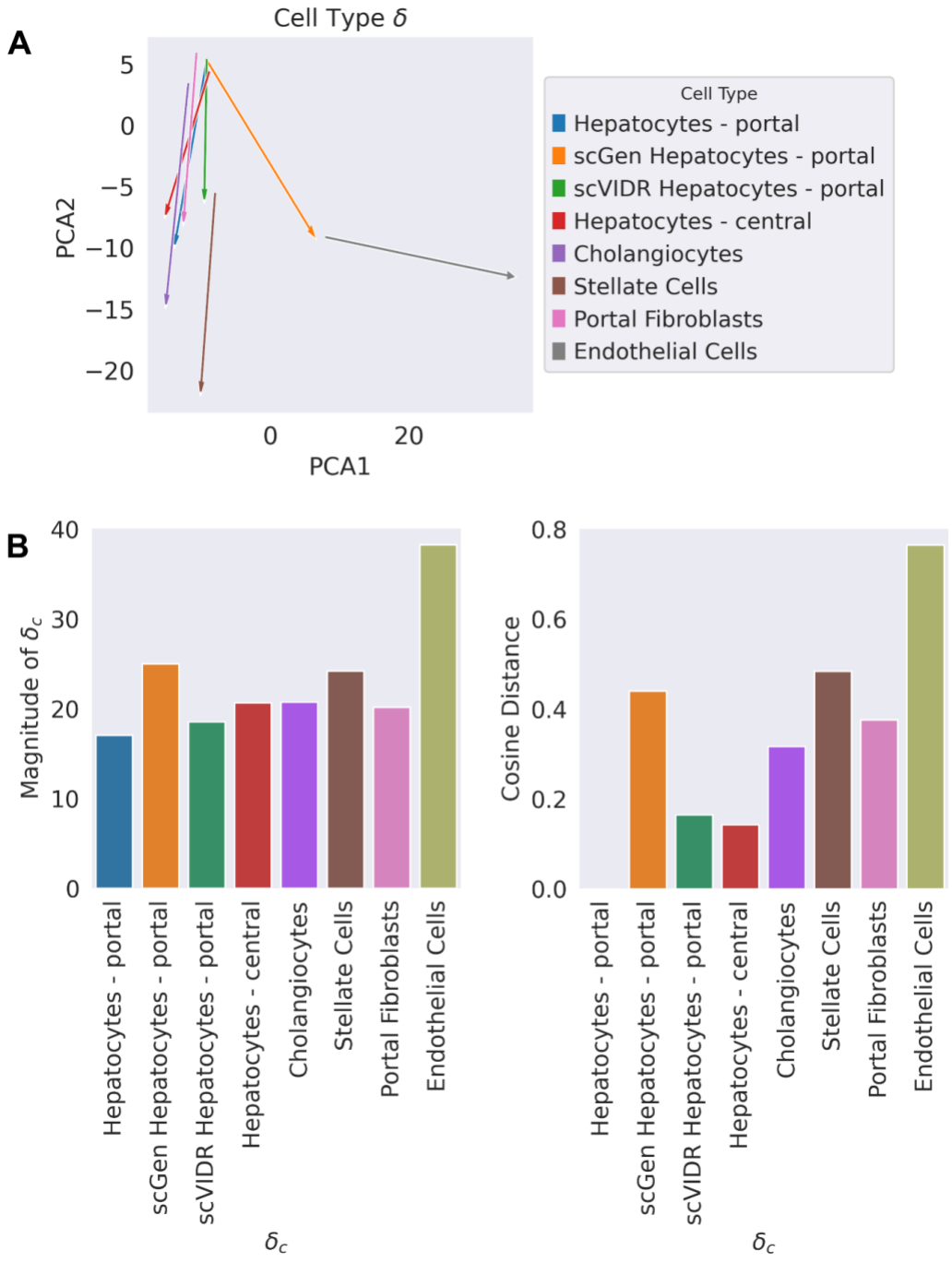
Supplemental information

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

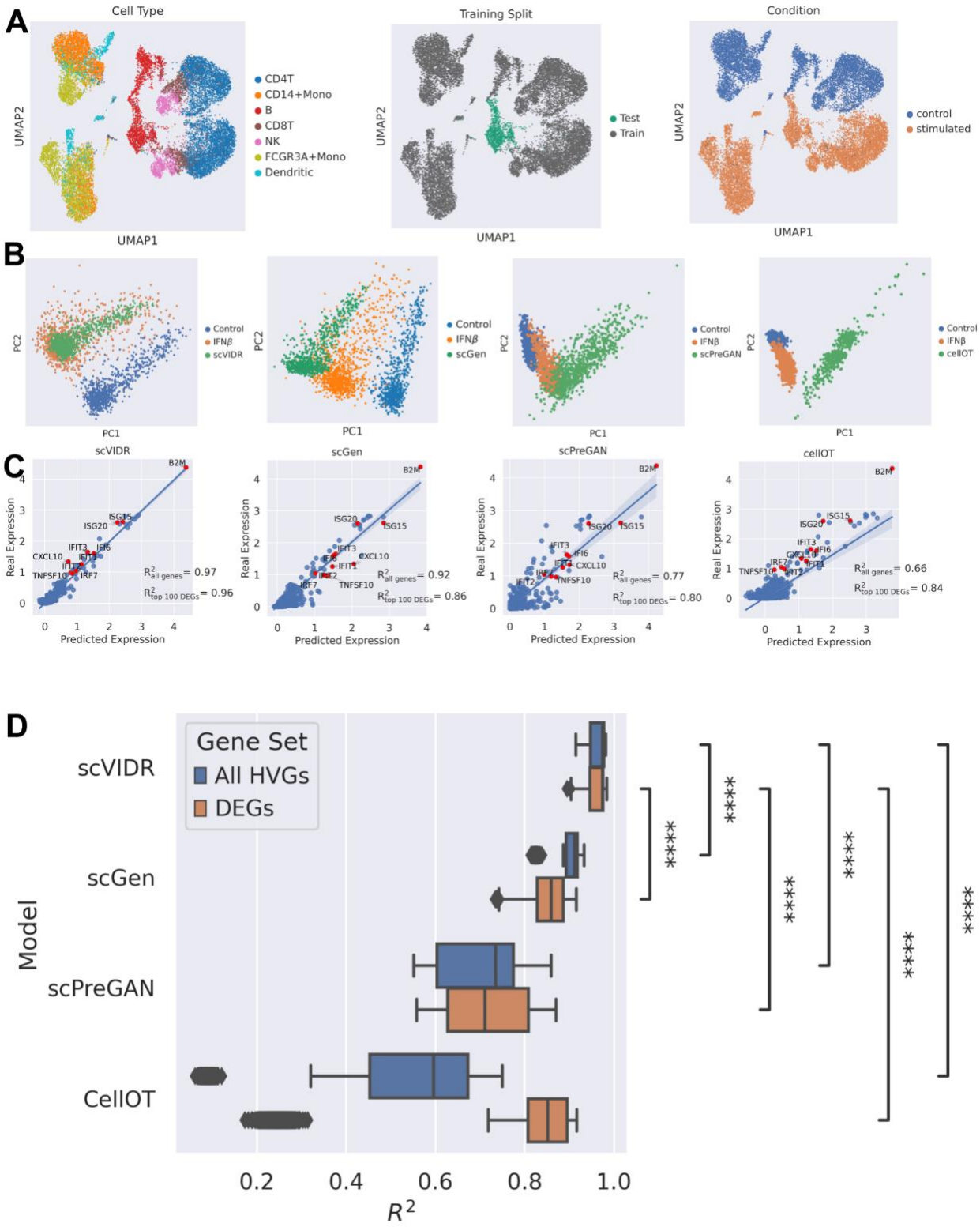
Omar Kana, Rance Nault, David Filipovic, Daniel Marri, Tim Zacharewski, and Sudin Bhattacharya



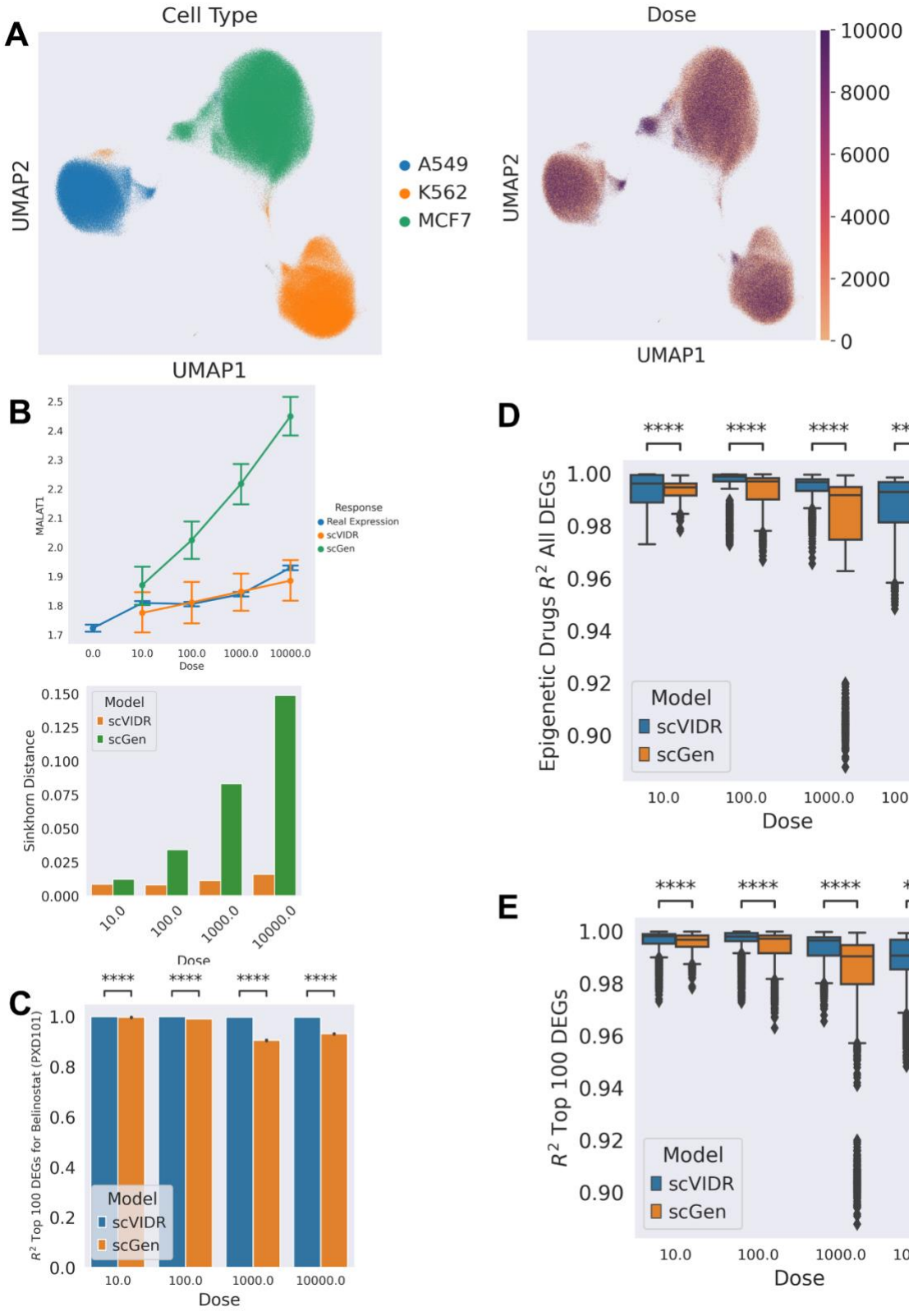
Supplementary Figure 1. Expression of AhR across liver lobule A) A zonal expression profile of normalized expression as described by Yang et al.[1] and Halpern et al.[2] Zone 0 represents the level of AhR expression in hepatocytes closest to the central vein. Zone 9 represents the level of AhR expression 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).



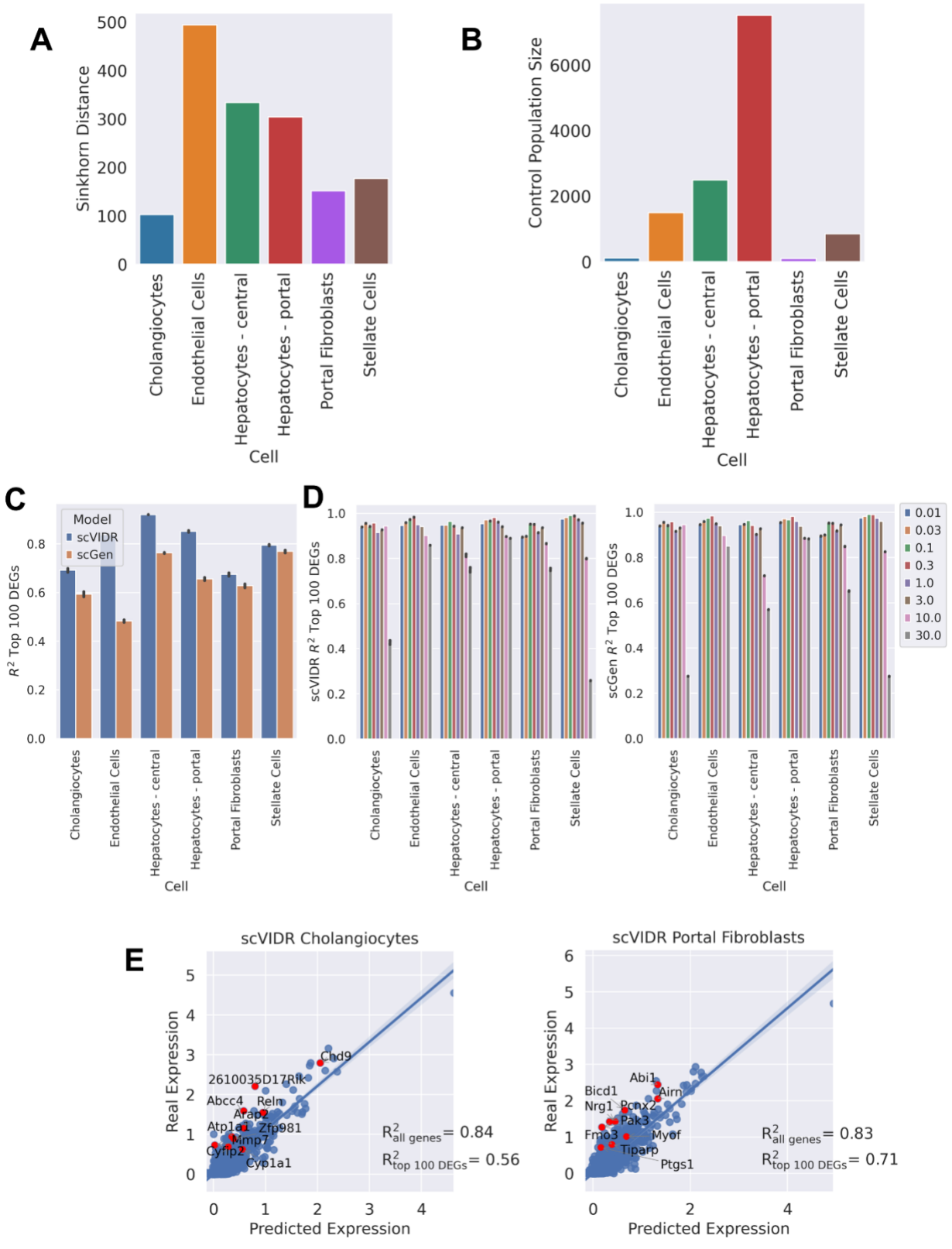
Supplementary Figure 2. δ_{scGen} deviates more from $\delta_{Hepatocytes-portal}$ than δ_{scVIDR} . A) A PCA visualization of the calculated δ_cs for a VAE trained without portal hepatocytes. "scGen Hepatocytes – portal" refers to the prediction by scGen (δ_{scGen})**,** and "scVIDR Hepatocytes – portal" 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
****

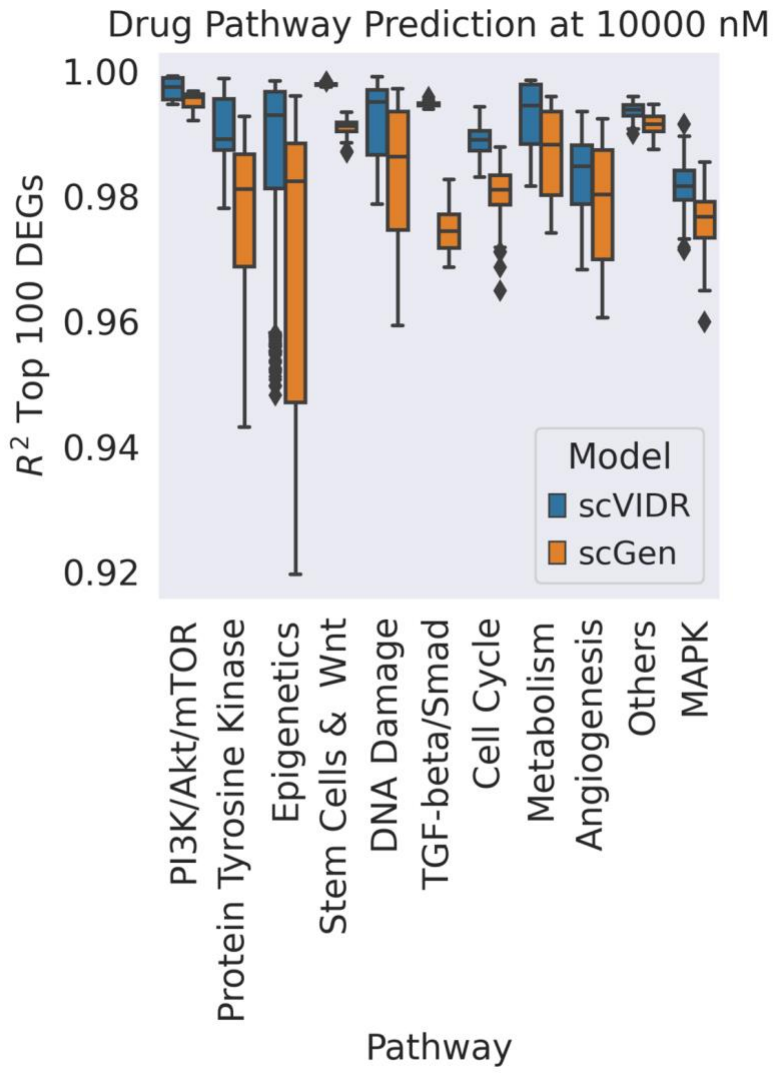
Supplementary Figure 3. Prediction of in vitro response of B-cells to IFNβ. A) UMAP of latent space of treated and untreated single-cell expression. UMAP plots are colored by cell type, training split, and condition, respectively**. B)** PCA plot of scGen, scVIDR, scPreGAN, and CellOT predictions of 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

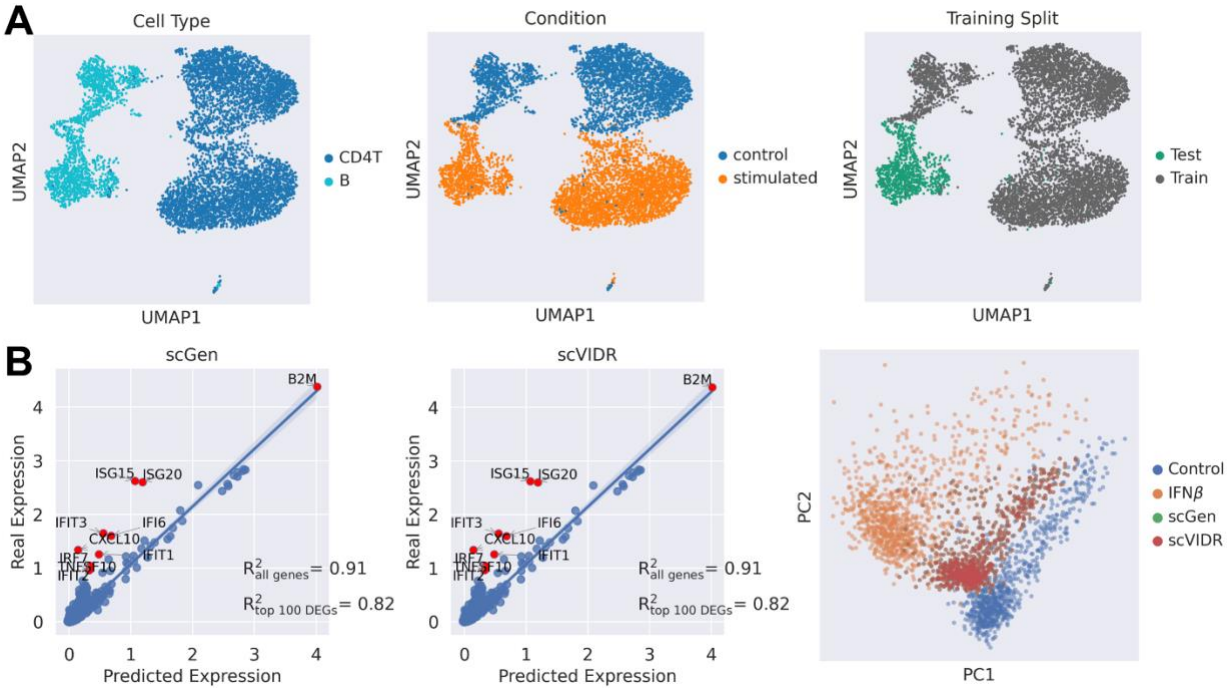
Supplementary Figure 4. Prediction of in vitro dose-response of A549 cells to different drug treatments. A) UMAP of the latent space of single-cell expression colored by cell type and dose (nM) respectively. **B)** Prediction of the dose-response of MALAT1 in response to Belinostat treatment of A549 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

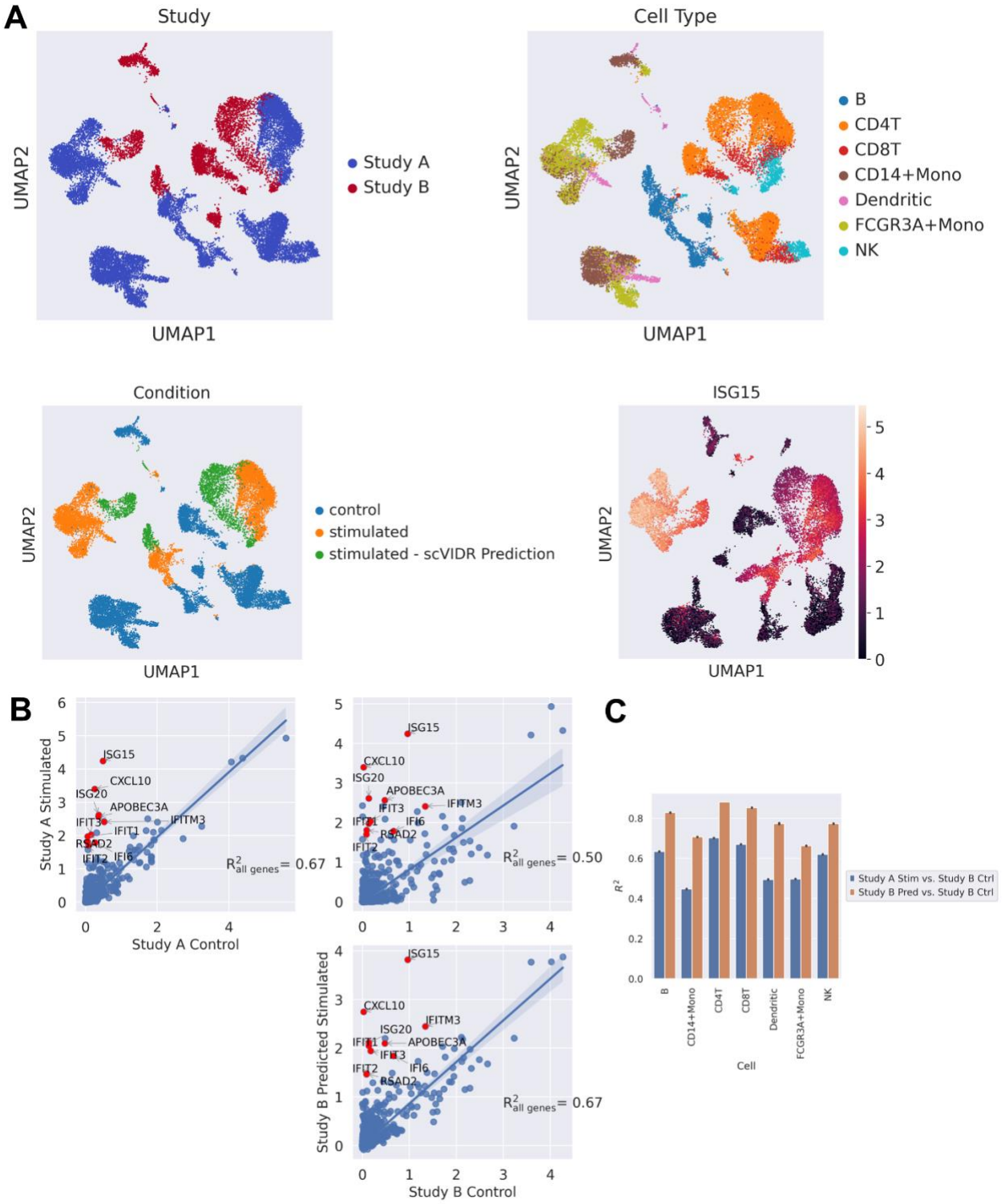
Supplementary Figure 5. Impact of latent perturbation magnitude and control population size on overall model performance. A) Sinkhorn distance between the latent distributions of the control and 30 μg/kg doses of TCDD of each cell type on the latent space. **B)** Bar plot of the control group cell population 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.



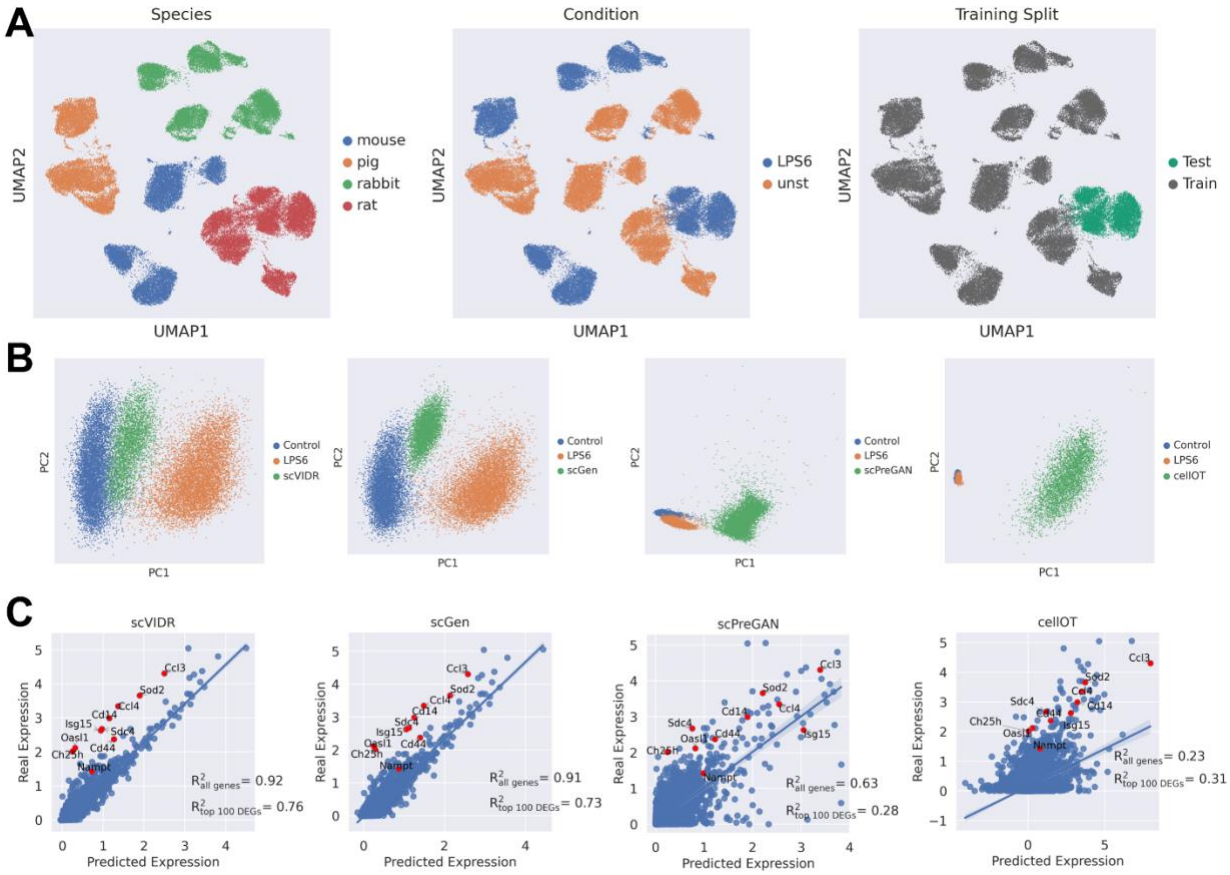
Supplementary Figure 6. Overall drug pathway performances at the highest administered dose in sci-plex dataset. A) A boxplot of the mean gene R^2 across all drug pathways in the test dataset at a dose of 10,000 nM.



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.