APPLICATION OF DEEP GENERATIVE MODELING TO SINGLE CELL RNA SEQUENCING DATA IN TOXICOLOGY

By

Omar Kana

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Pharmacology and Toxicology - Environmental Toxicology – Doctor of Philosophy

2024

ABSTRACT

Single cell RNA-sequencing provides an opportunity for pharmacologists and toxicologists study the heterogeneity of cellular responses to chemical perturbations. However, analyzing these datasets remains incredibly challenging as they are high volume and high dimensional. Efforts to mitigate these challenges have centered on dimensionality reduction. Recently, it has been shown that variational autoencoders (VAEs), deep generative models that can perform dimensionality reduction, can be successfully deployed on single cell RNA-seq data to perform normalization and prediction. In this thesis I will explore both use cases.

Even with the unprecedented detail single cell RNA-seq can provide for describing cell type specific chemical responses, exploring all relevant combinations of cell type-chemical perturbations remains difficult. Additionally, the dose of a chemical changes the overall character of its response. Variational autoencoders (VAEs) have been shown to predict chemical perturbations for single doses. However, VAEs have yet to be used to predict the entire dose-response. Here I introduce **single cell Variational Inference** of **Dose-Response** (*scVIDR*) which not only predicts the trajectory of the dose-response, but also achieves better large dose predictions than previous VAE algorithms. First, scVIDR is shown to predict dose-dependent gene expression across cell types in mouse liver, human blood cells, and cancer cell lines. Next, regression on scVIDR's latent space is used to biologically interpret model predictions. Finally, scVIDR is used to order individual cells based on their chemical sensitivity by assigning a pseudo-dose value. I conclude scVIDR can effectively be used to predict chemical perturbations in a wide range of administration scenarios.

Analysis of the compartmentalization of liver metabolism can be described with two axes: spatial and temporal. The spatial axis is conferred by the hepatic lobule, which is made up of concentric layers or zones of cells that have distinct metabolic programs. The master regulatory pathway that sustains this metabolic program is the Wnt/β-catenin pathway, which when activated by *Wnt* ligands, induces the transcription of potentiators of metabolic and nutritional gradients. The temporal axis of liver is conferred by the circadian rhythm which originates from super chiasmatic nucleus. Within each cell in the liver, this circadian rhythm is maintained by a core set of genes which confer feedback loops which sustain the necessary oscillations for metabolic efficiency. Previous studies have shown how the axes of the liver lobule can interact using single cell RNA-seq data. However, how these axes interact with toxicological perturbation is still

unknown. In this thesis, a variational autoencoder is used to batch correct single cell RNA-seq data for zonation inference. Existing models originally used to analyze the zonal-rhythmic axes of the liver lobule are extended to account for the effects of chemical perturbation. This methodology is applied to mouse liver snRNA-seq data from mice subjected to acute treatment of 2,3,7,8 tetrachlorodibenzo-*p*-dioxin.

This thesis is dedicated to the family who mean the world to me. To my mother, Rana, my father, Munib, my brother Yusef, and my wonderful fiancé, Lene. Thank you for loving and supporting me.

ACKNOWLEDGEMENTS

I cannot possibly acknowledge all who deserve it. I will try anyway.

I would first like to acknowledge my advisor, Dr. Sudin Bhattacharya, without whom my PhD would not have been possible. I have learned so much and gained such valuable experiences during my time in his lab. I thank him for creating the space I needed to develop as a scientist.

I would like to acknowledge my committee members: Dr. Cheryl Rockwell, Dr. Norbert Kaminski, and Dr. Rance Nault. All of whom have been generous with their knowledge and data. Dr. Nault specifically has given me such wonderful opportunities and insight that I believe without his being there, my PhD would have been incredibly difficult.

I would like to acknowledge the members of my lab: Dr. David Filipovic, Mr. Daniel Marri and Ms. Leah Terrian. All of whom are like friends to me. Their support and companionship over the last five years has eased many a stressful situation. To David, I would like to extend a personal thanks, as your insight and counsel has had a direct impact not only on my scientific career, but also on my outlook on life.

I would like to acknowledge my best friends, Jason, and Tofiq, who have stuck with me since high school. Returning home to Louisiana during COVID-19 epidemic, and those brief visits during holidays were always richer when you were there to listen to my ramblings about graduate school.

I would like to acknowledge my family. My mother, father, and brother have only ever been supportive since I started my PhD. Your love is like the solid ground beneath my feet, supporting me wherever I go. You have always been the driving force pushing me forward to be my best.

Finally, I would like to acknowledge my fiancé, Lene, whose presence in my life over the past two years has been wonderful. You have made every challenge easier, every stressful situation comfortable, and every cloudy Michigan day brighter. I am so thankful for your patience and support during my PhD.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AhR - Aryl Hydrocarbon Receptor

ALT – Alanine Transferase

AOP – Adverse Outcome Pathway

BIC – Bayesian Information Criterion

ChIP – Chromatin Immunoprecipitation

CITE-seq – Cellular Indexing of Transcriptomes and Epitopes sequencing

DEG – Differentially Expressed Genes

δ – The mean difference between the treated and control populations

GM-CSF – Granulocyte-Monocyte Colony Stimulating Factor

GSEA – Gene Set Enrichment Analysis

HVG – Highly Variable Gene

INF-β – Interferon Beta

LPS – Lipopolysaccharide

MMD – Maximum Mean Discrepancy

MNLEM – Mixed Non-Linear Effects Model

NAFLD – Non-alcoholic Fatty Liver Disease

DILI – Drug induced liver injury.

PBMC – Peripheral Blood Mononuclear Cell

PC – Principal Component

PCA – Principal Component Analysis

scVIDR – single cell Variational Inference of the Dose Response

scATAC-seq – Single Cell assay for transposase-accessible chromatin

scRNA-seq – Single Cell RNA sequencing

snRNA-seq – Single Nuclei RNA sequencing

TCDD – 2,3,7,8 Tetrachlorodibenzo-*p*-dioxin

UMAP – Uniform Manifold Approximation and Projection

VAE – Variational Auto-encoder

ZT – Zeitgeber Time

CHAPTER 1
INTRODUCTION

1.1 The combinatorial problem in pharmacology and toxicology

Drug discovery presents a difficult problem to scientists. The space of possible biochemical therapeutics is extremely large[1]. Just for small organic chemicals, there exist 10^{60} possible structures, and of those potentially 10^{23} structures could be pharmacologically relevant[2]. This space of chemicals does not even include the biologics and inorganic compounds that exhibit therapeutic properties. Despite this large chemical space for drug discovery, many drugs will not make it to market. This is due to a myriad of reasons, including, prominently, potential toxicological effects on any different number of physiological systems[3].

Physiological responses to the same chemical can vary from cell type to cell type[4], and according to Tabula Sapiens[5], a molecular reference atlas of human cells, there are at least 400 distinct human cell types. Within those cell types, there exist potential physiological gradients and differences in function based on the context. An example explored in this thesis is the metabolic gradient in the liver lobule, where hepatocytes nearer to the central vein and those closer to the portal triad perform different and complementary sets of metabolic functions (e.g., Ahr expression; Figure 1.1)[6]. In addition to these metabolic gradients, stochasticity resulting from the biochemical process of transcription and translation, global differences in cellular parameters (e.g., gene copy number), and the chemical environment of the cell result in variation in the responses of cells of the same type[7–9]. As a result, the spectrum of cellular response to chemical perturbation gives us an incomprehensibly large combinatorial space that must be explored by toxicologists to determine chemical safety.

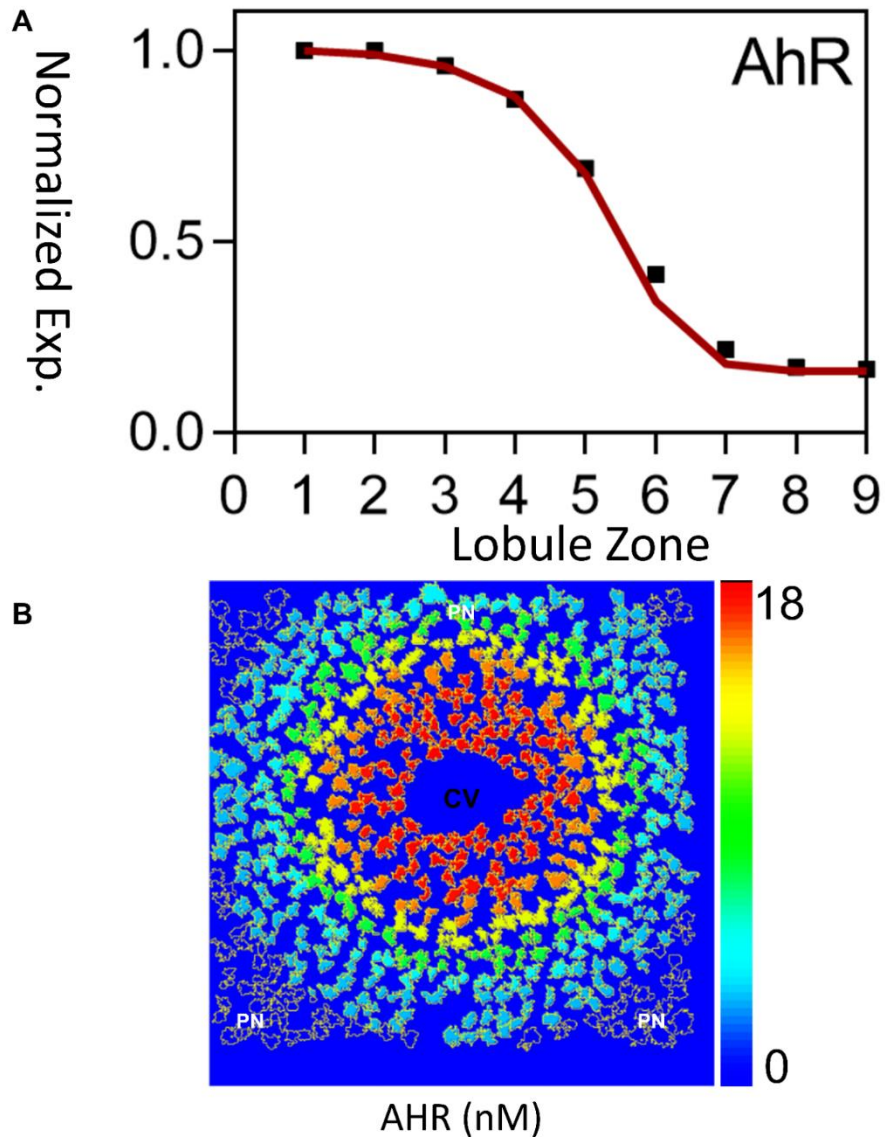

Figure 1.1 Expression of AhR across hepatic lobule. A) A zonal expression profile of normalized expression as described by Yang et al[10] and Halpern et al[6]. Zone 0 represents the level of AhR expression in hepatocytes closest to the central vein. Zone 9 represents the level of AhR expression closest to the portal vein. B) A single-cell resolution image of the liver lobule generated by Halpern et al.[6] with AhR expression levels represented by color from Yang et al.[10] The central vein is denoted by "CV" (label in black) with the portal triad denoted by "PN" (label in white).

In the past, this problem was not addressed due to the experimental limitations in measuring the state of a single cell. However, recent decades have created an omics revolution, in which

scientists can profile thousands of different endpoints from protein[11] and RNA expression[12] to the presence and concentration of particular metabolites in a cell[13]. This thesis will mainly focus on data from single cell RNA sequencing (scRNA-seq)[14–16], which can be used to profile the transcriptomes of tens of thousands of individual cells at once. This presents an entirely new level of resolution for toxicologists, as for the first time ever, the true heterogeneity of cellular transcriptomic response can be profiled in an efficient manner[17]. My thesis presents a method to analyze and predict high dimensional gene expression data and heterogeneity in cellular dose-response.

1.2 The unreasonable effectiveness of dimensionality reduction

In the statistical field of machine learning, there is a class of algorithms that perform a task called dimensionality reduction. The analysis and visualization of high dimensional data like single-cell gene expression can be facilitated by a broad class of dimensionality reduction algorithms. Examples of such algorithms include principal component analysis (PCA)[18], which aims to preserve the global variances of the data, and uniform manifold approximation and projection (UMAP)[19] which aims to preserve the local distances in low dimensions (Figure 1.2 B and C). This class of algorithms is routinely used in the study of scRNA-seq data and is useful for the visualization of high-dimensional datasets[20], as well as in downstream tasks such as clustering and trajectory analysis[21].

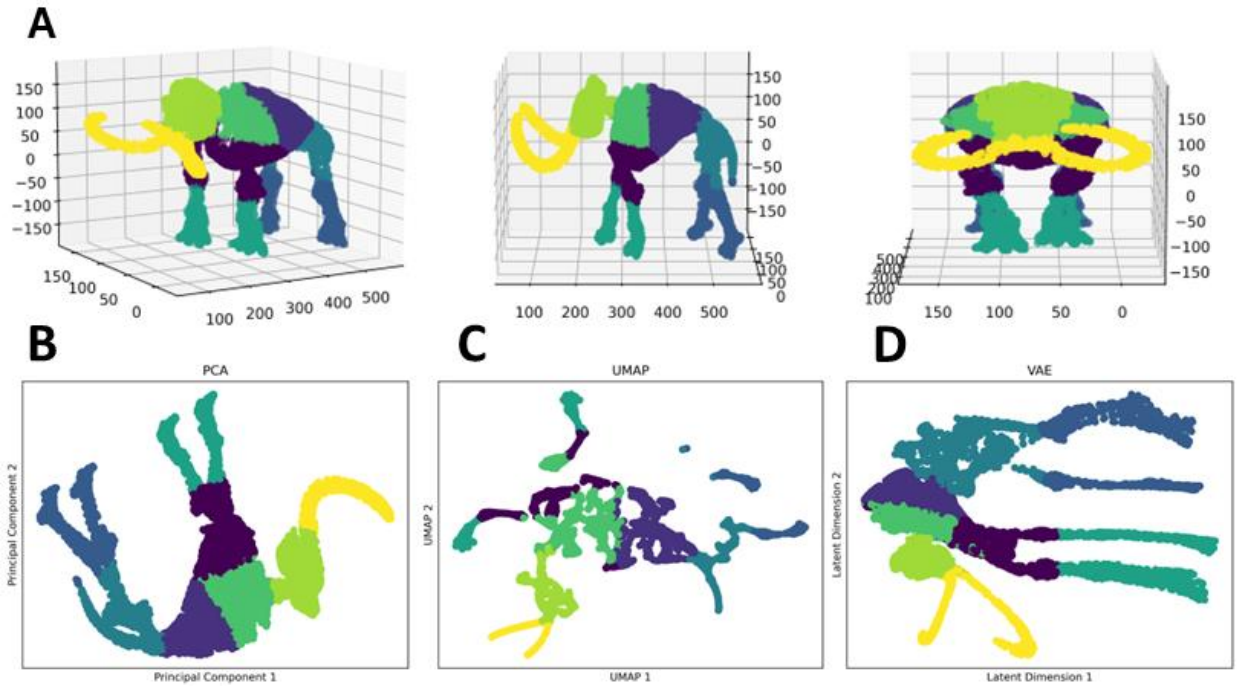

Figure 1.2 Dimensionality reduction of 3D mammoth. A) Point representation of wooly mammoth (Smithsonian Institution Archives USNM: V23792) at three different perspectives in 3D space. Colored using K-means clustering. B) 2D representation of wooly mammoth produced by principal component analysis (PCA). PCA preserves inter-cluster distances better than intra-cluster distances. C) 2D representation of wooly mammoth produced by Uniform manifold approximation and projection (UMAP). UMAP preserves intra-cluster distances better than inter-cluster distances. D) 2D representation of wooly mammoth produced by Variational autoencoder (VAE). VAE's utilize neural networks to do dimensionality reduction and optimize both inter- and intra-cluster distances.

Dimensionality reduction is remarkably efficient, often needing only orders of magnitude fewer features to properly represent a space. This can be seen even in large atlas sized scRNA-seq datasets made up of hundreds of thousands of cells. An example of such a dataset is the Tabula Sapiens[5] which contains ~500,000 cells across 24 tissues from 15 human donors. Despite having measurements for ~30,000 genes, 100 principal components explain ~40% of the variance (Figure 1.3). This efficiency is reflected in the ubiquity of PCA's use as it is a common preprocessing step in scRNA-seq pipelines to use the first 50-100 principal components[22]. With thousands of possible measurements in hundreds of thousands of cells, it is striking that with

only 50-100 components, one can capture significant variation in the dataset. A possible explanation lies in the manifold hypothesis[23]. Essentially, while real-world datasets can reside in high-dimensional space, they can often be described with a local-coordinate system of fewer dimensions. These dimensions represent a more concise summary of the high-dimensional space and thus make it more useful for experimentalists to describe a system. This is likely to be true in the case of cell biology, where gene-gene regulatory interactions constrain the number of possible attainable cell states[24,25]. This is particularly helpful in the domains of pharmacology and toxicology. While there are potentially 10^{23} pharmacological drugs[2], chemicals with similar structures will tend to induce similar effects in the same biological system[26]. Due to the large number of genes in the genome, there is an infinite set of configurations of cellular state. However, the interactions between genes and the process of evolution limit the possible range of phenotypes to a much more tractable amount[27]. Which is why while omics can provide measurements of 1000s of dimensions, they are amenable to dimensionality reduction.

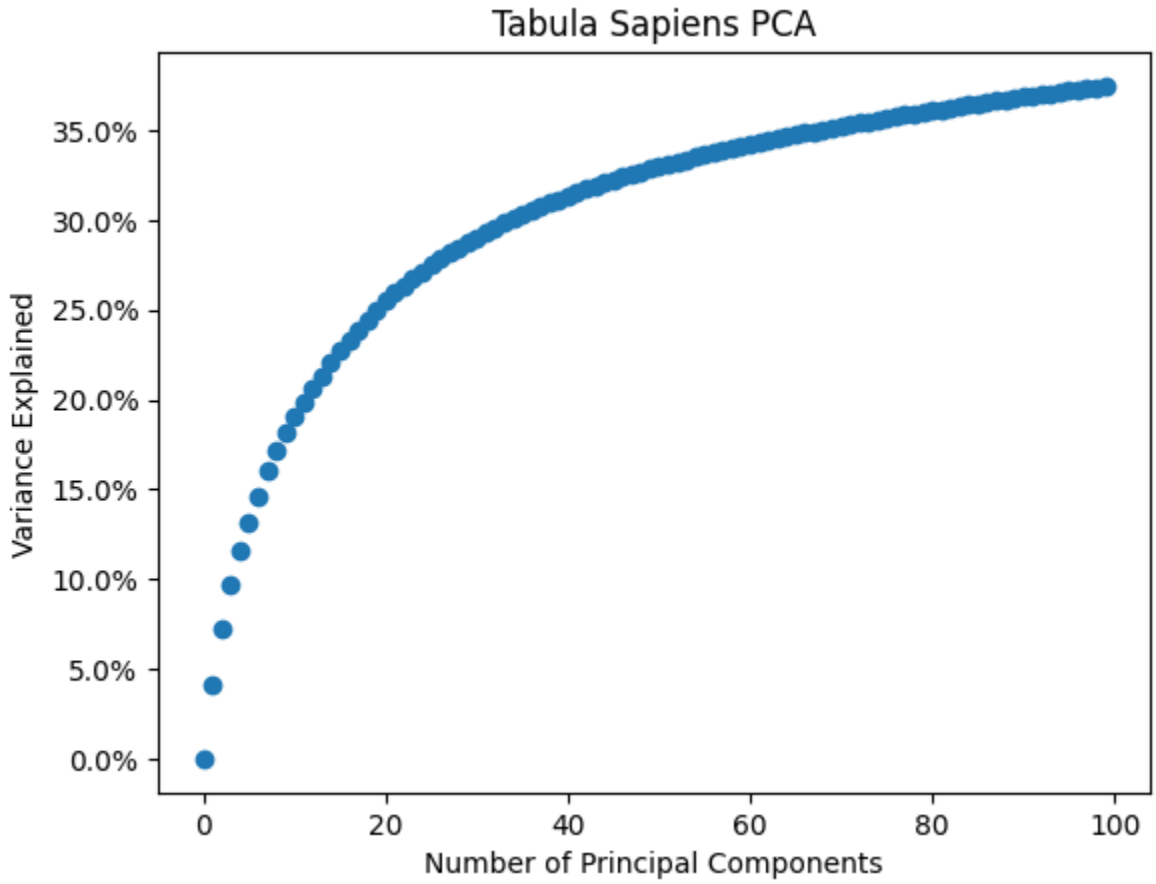

Figure 1.3 Cumulative explained variance in PCA of Tabula Sapiens. PCA done with differing number of components was performed on the Tabula Sapiens. The x-axis is the number of principal components used in the dimensionality reduction. The y-axis is the resulting variance ratio explained by the PCA.

For the purposes of this thesis, I choose to focus on dimensionality reduction using a class of deep neural networks called variational auto encoders (VAEs)[28] (Figure 1.2 D) which have been used extensively to model and make predictions of changes in single cell gene expression[29–34].

1.3 A Brief Introduction to Variational Autoencoders

VAEs are a class of deep generative models which rely on Bayesian priors to encode single cell data into a latent distribution[28]. VAE's represent a subclass of autoencoders, which themselves are a generalization of PCA[35]. VAEs can be seen as a non-linear "cousin" to PCA, with certain advantages and disadvantages when comparing the two techniques. To introduce the reader to VAEs and how they work, I will first describe the relationship between PCA and autoencoders to

frame the discussion. I will then describe the basic structure of an autoencoder. Finally, I will describe how the autoencoder framework is extended to variational autoencoders. A full mathematical description of variational autoencoders can be found in section 3.4.1.

As described in section 1.2, PCA is a form of dimensionality reduction that is used extensively in scRNA-seq data. PCA is a linear transformation of the data which aims to preserve as much variance as possible for each dimension[18]. A forward function brings the high dimensional scRNA-seq measurements to a lower dimensional representation. This lower dimensional representation, also known as principal component (PC) space, can be used for downstream analysis such as visualization [36], clustering of the data for cell type identification [37], and input into trajectory inference algorithms (i.e., pseudo-time)[21]. An advantage of PCA is that since it is linear the user can directly interpret the dimensions of the lower dimensional space using loadings[38,39]. These loadings can be used to describe gene expression programs in each principal component of the lower dimensional space. An additional useful feature of PCA is that the forward function is invertible, and the input data can be approximately reconstructed (Figure 1.4 A). However, trying to generate novel snRNA-seq measurements not included in training will often result in inaccurate reconstructions[30]. To address this problem, non-linear functions can replace the forward and inverse functions.

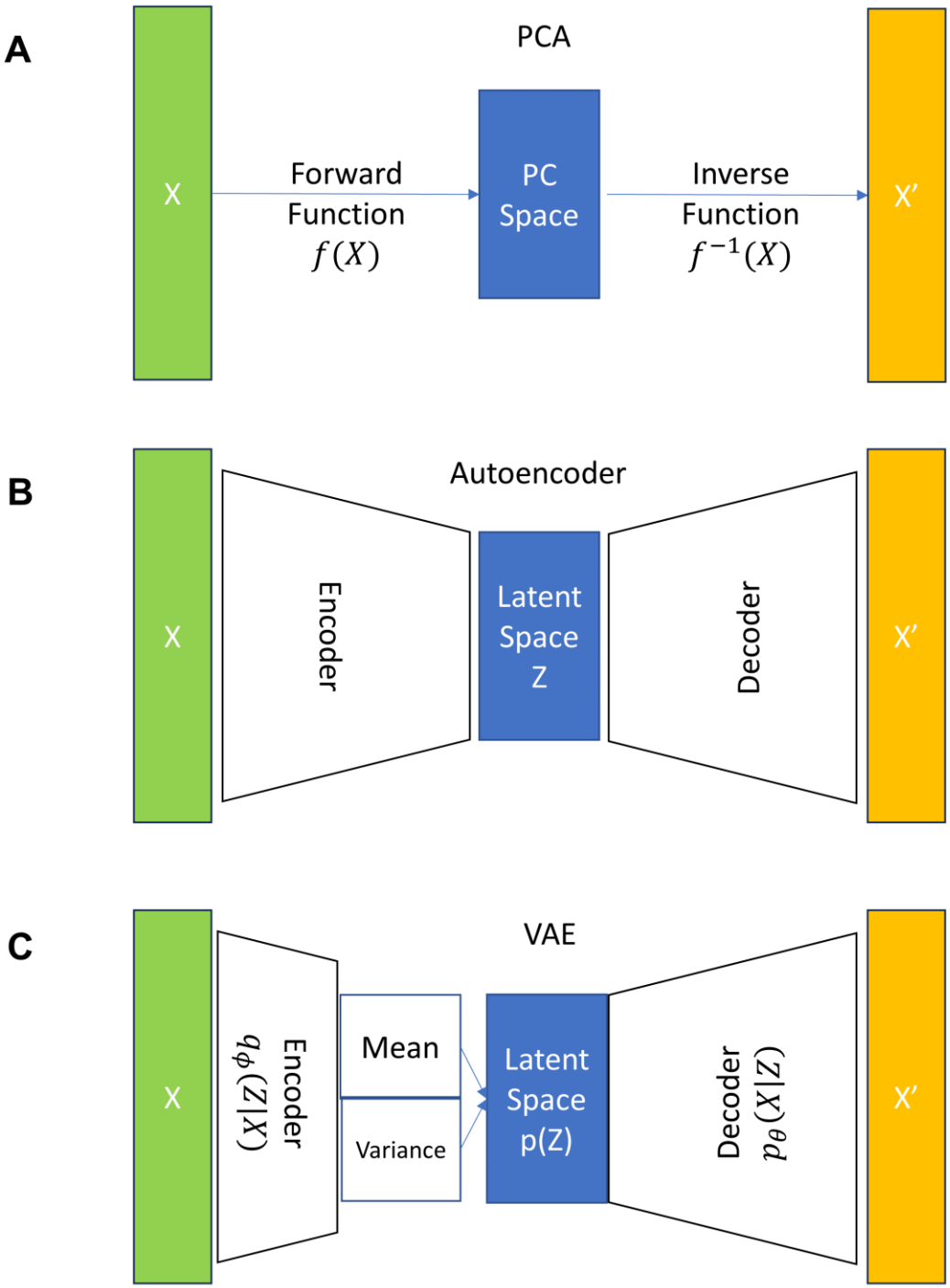

Figure 1.4 Diagrams of PCA, an Autoencoder, and a VAE. Input scRNA-seq measurement data is represented by the green box labeled X and the reconstructed scRNA-seq data from the lower dimensional representation is represented by the orange box labeled X'. Lower dimensional representations are shown as blue boxes. A) Schematic of PCA. scRNA-seq measure data is orthogonally transformed using a linear $f(X)$ into a lower dimensional principal

Figure 1.4 (cont'd)

component (PC) space. The inverse function, $f^{-1}(X)$, is used to take points from PC space back into scRNA-seq measurement space. B) Schematic of an autoencoder. Autoencoders are like PCA, but instead of linear transformations f and f^{-1}, there are non-linear neural networks that replace them called an encoder and decoder respectively. The encoder takes the scRNA-seq measurement, and compresses it into the latent space, Z. The decoder takes points from Z and approximately reconstructs them back into scRNA-seq measurement space. C) Schematic of a VAE. Unlike a normal autoencoder, a variational autoencoder has a probabilistic encoder and decoder. A normal VAE encoder, $q_\phi(Z|X)$, outputs a mean and variance which is used to sample the features of the latent space, $p(Z)$. The decoder then approximately maps the probability distribution of the latent space back into scRNA-seq measurement space.

Autoencoders aim to perform the dimensionality reduction using deep learning. Autoencoders replace linear forward function and inverted function in PCA, with neural networks[35]. The neural network replacing the forward function, which compresses the high dimensional scRNA-seq data to a lower dimensional representation, is called the encoder. The lower dimensional representation of the scRNA-seq is referred to as the latent space. The neural network replacing the inverted function, taking latent space measurements back into scRNA-seq measurement space, is called the decoder (Figure 1.4 B). The use of neural networks instead of linear functions is useful, since now the user can calculate more accurate lower dimensional representations of the scRNA-seq measurements[40,41]. Accurate representations are critical during scRNA-seq processing pipeline, as the cell type identification step requires a lower dimensional representation of the data as input for clustering[42]. This comes at a cost, however, as the non-linearity of the neural networks makes interpretation of the latent dimensions more difficult as there are no longer any linear loadings. In section 4.2.3 and section 4.4.4 of this dissertation, I describe a method to overcome this limitation in VAEs and interpret them using a loadings like model. Additionally, the autoencoder's latent space dimensions are not forced to be structured[43], meaning that dimensions of the latent space are entangled with one another and are not guaranteed to be independent of one another like in PCA. As a result, vanilla autoencoders have latent spaces that are much less generalizable to other datasets[35].

The variational autoencoder represents an extension of autoencoders that impose more structure on the latent space using Bayesian priors. This imposed structure forces the latent dimensions to be as disentangled as possible. To do this, instead of the dimensionality reduction being a deterministic function like in vanilla autoencoders, variational autoencoders describe the data generation process probabilistically[35]. This means that instead of encoding a lower representation directly, the encoder calculates a mean and variance of a gaussian distribution, from which the user can sample features for the latent space (Figure 1.4 C). This gaussian distribution is called the prior distribution and is what imposes structure one the latent space. To ensure that the encoder's output is as close to prior (i.e., structured) as possible, the neural network minimizes the difference between the encoder output and the prior distribution. This imposed structure in the latent space is what allows the VAE to generate non-random samples from the latent space. In practice this means that the decoder can generate meaningful scRNA-seq measurements from unobserved points in the latent space rather than just the points in the training data. Thus, VAEs have the non-linear power of autoencoders with added generalizability to other datasets[35]. Due to these advantages, VAE's have proven to be an incredibly flexible and effective tool in the analysis of scRNA-seq[30,31,33,34,44–47].

1.4 Variational autoencoders and single cell RNA sequencing

VAEs have been used extensively across a spectrum computational single cell task. At its most basic, VAEs have been used to visualize scRNA-seq data[44,45], and at its most ambitious, has been used to integrate scRNA-seq data with other single cell data modalities[34]. Most single cell research on VAE's has been in correcting scRNA-seq data for batch effects[29]. However, VAE's have extended use in other problem domains such as cell type identification[46], integration of datasets from multiple labs[48], and inference of cell-cell interactions[49]. The following section is a brief survey of the current field, with a focus on batch correction, cell type identification, and prediction of chemical perturbations.

VAEs have been used to integrate single cell RNA-seq across experimental groups[29], labs[48], and data modalities[34]. Integration, also called batch correction, is a process by which the variation of the data based on some batch covariate (e.g., sex, age, lab of origin, etc.) is removed from the data. This is often done in order to properly address whether the changes in the data originate from the variable of interest (i.e., TCDD treatment) independent of these batch effects. VAE models such as scVI have been used to model such covariates including library size and single

cell chemistry[29]. To do this, models will condition the encoder and decoder on the batch labels in order to remove their variance from the latent space. For integrating across different labs of origin, models such as scArches[33,48] further force integration by adding a regularization term to the loss function. This term causes the model to bring the batch clusters as close together as possible on the latent space. VAE models such as MultiVI have been used to integrate multiple single cell data modalities (i.e., snRNA-seq, CITE-seq, and single cell ATAC-seq)[34]. To do this, MultiVI first encodes each modality into its own latent space. Then MultiVI averages the latent spaces together to make the final single latent representation of all data modalities. To make sure that the different latent spaces can be averaged together, an additional term is added to the loss function. This term forces the different latent spaces to be as similar to one another as possible. As with the normal integration, the aim is to make an integrated representation across all data modalities such that one can compare impacts across chromatin accessibility, protein expression, and transcription. In chapter 2 of this thesis, I describe the use of variational autoencoders to model and remove effects of TCDD on hepatocytes to infer liver zonation more accurately from scRNA-seq data.

Conditional latent representations have been used in cell type identification. To do this, models will utilize existing large datasets of tissues such as the Allen Mouse Brain Atlas[50], or the Tabula Sapiens[5]. These models use these large tissue datasets as reference to train the VAE's, after which the VAE will then classify cells in new datasets based on their position in the latent space. Additionally, since these models can integrate across labs of origin, the models can utilize multiple large cell atlases to classify new cells. One approach, pioneered by scANVI[46], is to utilize existing architecture created by scVI[29] and extend it by incorporating a neural network classifier. The classifier and the VAE are trained in parallel so that the latent space is optimized to make it as simple as possible for the neural network to distinguish between different classes of cells. Alternative models for cell type classification include MoE-Sim-VAE which utilizes a mixture of experts architecture to cluster and identify different cell populations[51].

One pharmacologically focused application of variational autoencoders for the prediction of chemical perturbations[30,33]. In section 1.3, one of the mentioned advantages of VAE's were that they were generative models, and as a result could generate meaningful scRNA-seq measurements from unobserved points in the latent space. Recently, VAEs have been demonstrated to predict the expression of a chemical perturbation on an unobserved cell type

based on the perturbations of other cell types[30]. To understand how this might work, I take an example of the model in another discipline, computer vision. Much like scRNA-seq, images are high dimensional data (each pixel representing a dimension) which can be compressed into a lower dimensional space using variational autoencoders. Let us consider a picture of a woman's face, to which I want to virtually add glasses. This can be accomplished using VAEs by subtracting the latent representation of pictures of men's faces from those of men's faces with glasses, and subsequently adding the resulting "glasses" vector to pictures of women's faces without glasses.

I can imagine an analogue to this model, but instead of gender I have a cell type and instead of glasses I have a chemical perturbation. These models are examples of using vector arithmetic on the latent space of a VAE. One such model, scGen[30], has been shown to outperform other generative models such as generative adversarial networks and other dimensionality reduction algorithms such as PCA on prediction or chemical perturbations for unseen cell types[30]. However, these models have difficulty accounting for certain complexities in biological data. For one, the response of a particular drug is highly dependent on cell type[8,9]. Thus, a simple addition of a perturbation must be weighted on the cell's transcriptomic profile. Additionally, the magnitude of the chemical perturbation must also be considered. At lower concentrations, there may be little to no effect of a drug on a cell's transcriptome. However, at higher concentrations, these effects can be more pronounced. Furthermore, even if I can make these predictions, interpretation of how the model makes predictions on a gene-by-gene basis must also be considered to evaluate whether the model is describing the chemical perturbation appropriately. In chapter 3 of this thesis, I improve on the original vector arithmetic method using regression on the latent space. In chapter 4, I further extend the model to account for multiple doses, interpret prediction on a gene-by-gene basis, and order cells based on how perturbed their transcriptomes are.

CHAPTER 2
PERTURBATION OF LIVER RHYTHMICITY AND ZONATION BY 2,3,7,8 TETRACHLORODIBENZO-P-DIOXIN

2.1 Introduction

To deal with the multitude of chemicals that an organism will need to interact with on a regular basis and to sustain healthy metabolic homeostasis, the liver has evolved to compartmentalize metabolic programs[52]. This compartmentalization is both spatially and temporally organized. Spatially organized via the porto-central axis of the hepatic lobule, repeating hexagonal sub-units of the liver[53]. Temporally organized via the circadian rhythm based on feeding/fasting cycles[54,55]. In previous studies, the spatial and temporal metabolic organization of hepatocytes was described for over five thousand genes in hepatocytes[52]. However, it is unknown how resilient hepatic compartmentalization is to acute toxicological perturbation. In this chapter I investigate changes in the spatial and temporal axes of the liver with respect to acute treatment of 2,3,7,8 tetrachlorodibenzo-*p*-dioxin (TCDD). To do this I first infer the zonation of the liver lobule using a Variational autoencoder[28,29] and diffusion maps[56,57]. Then I extend a method introduced by Droin and Kholtei et al[52] to classify genes based on their rhythmicity and zonation to include classifications of TCDD influence.

The histological unit of the liver is the hepatic lobule (Figure 2.1). The hepatic lobule is a hexagonally shaped structure with a central vein at the center and a portal triad at each vertex of the hexagon[6,53]. Each portal triad is made up of a portal artery, portal vein, and a bile duct[58]. Due to the positioning of vasculature and the direction of blood flow (from portal to central), nutrient and metabolic gradients are established within each lobule that confer different metabolic functions at different radii extending from the central vein[53]. For example, hepatocytes nearer to the central vein will tend to have higher expression of pathways involved in metabolism of xenobiotics by CYP450 genes[59]. Hepatocytes nearer to the portal triad will tend to specialize in β-oxidation and gluconeogenesis[52]. This metabolic gradation over the axis from the portal triads to the central vein (porto-central) is called liver zonation[6,52,53,60]. Liver zonation is maintained by a complex network of chemical cues and cell-cell interactions[60]. The canonical master regulatory pathway is the *Wnt*/β*-catenin* pathway[61]. In brief, *Wnt* proteins bind to *Frizzled* causing lipoprotein receptor-related proteins to phosphorylate the β-catenin degradation complex. This causes β-catenin to dissociate from the degradation complex, translocate to the nucleus, and activate potentiators of zonation[62,63]. Traditionally, liver zonation refers to the categorization of

hepatocytes either residing the pericentral (referred to as central; zone 3), mid-lobular (zone 2), and periportal (referred to as portal; zone 1)[53]. However, in more recent studies zonation suggested a finer grained, or continuous pattern of concentric layers of hepatocytes[6,52].

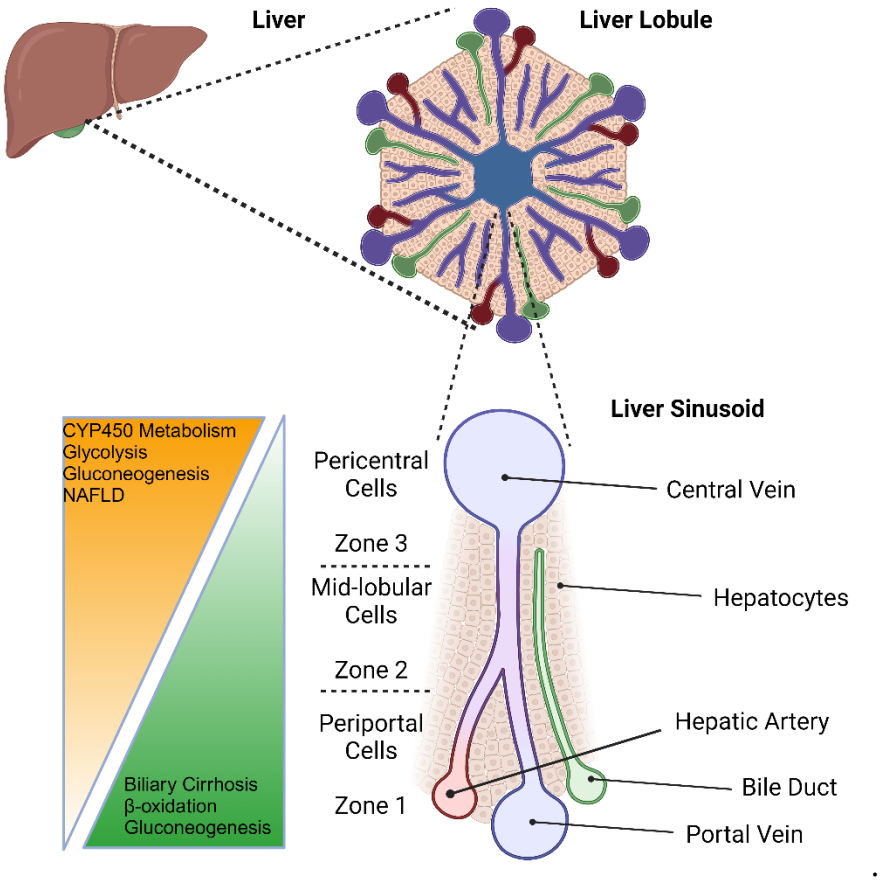

Figure 2.1 Cartoon of hepatic zonation within the liver lobule. Within the liver there exist hexagonal hepatic subunits known as liver lobules. Within these lobules exist a gradient of nutrient, oxygen, and metabolic functions along each liver sinusoid from the central vein to the portal triad (i.e., bile duct, portal vein, and portal artery). Cells along the sinusoid are separated into three zones (periportal zone 1, mid-lobular zone 2, and pericentral zone 3). The metabolic gradient of periportal functions and pathologies are represented by a green triangular gradient. Metabolic gradients of pericentral functions and pathologies are represented by an orange triangular gradient.

Temporal regulation of the liver is described using the circadian rhythm. Temporal compartmentalization of metabolic functions makes sure that liver function (e.g.,

glycolysis/gluconeogenesis) is coordinated with feeding/fasting cycles of the organism[52,54,64–66]. Disruptions of the circadian rhythm are associated with diseases such as non-alcoholic fatty liver disease (NAFLD)[66–68]. Thus, sustaining a consistent rhythm within the liver is important to the overall health of the organism.

Within each cell is a molecular oscillator that is made up of a network of feedback loops. In hepatocytes, this oscillator is made up of a core set of circadian genes (Figure 2.2)[69]. CLOCK and ARNTL bind to one another forming the CLOCK-ARNTL complex. The CLOCK-ARNTL complex then binds to E-box motifs which activate the transcription of downstream core circadian genes. These genes include genes that translate to the PER-CRY complex which inhibits the binding of CLOCK-ARNTL. CLOCK-ARNTL also activates genes that translate to REV-ERB and DBP. DBP upregulates *Nr1d1/2* (genes that encode REV-ERB) which translates to more REV-ERB. REV-ERB competes with ROR transcription factors and inhibits the translation of further *Nr1d1/2* and *Arntl*. These feedback loops make the basis for the rhythmicity seen in hepatocytes[69].

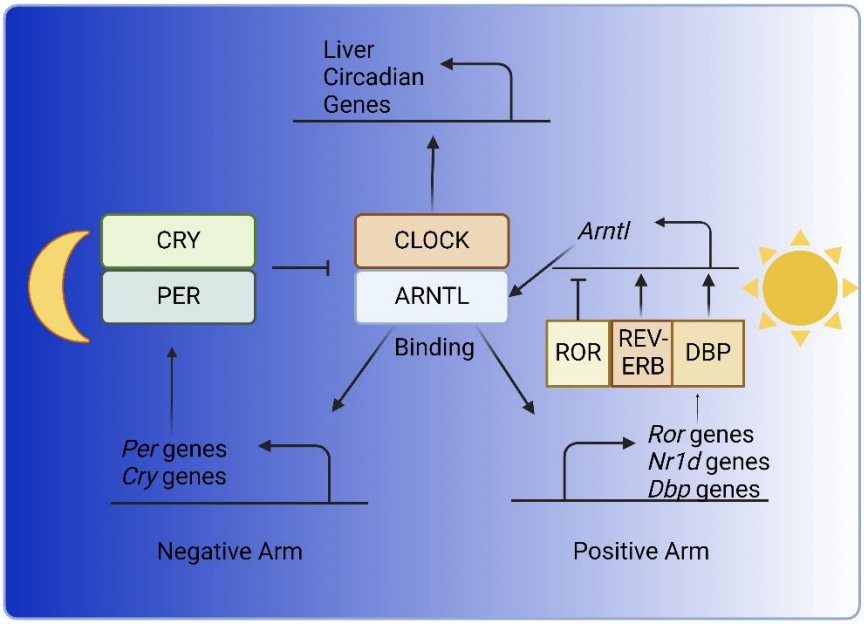

Figure 2.2 Molecular pathway of circadian rhythm in liver. Day-night cycle represented with box containing blue gradient. Proteins like CLOCK and DBP are indicated by colored boxes within the day-night diagram. Transcription of genes represented by curved arrows. Activation is represented by arrows. Inhibition is represented by flat-head arrows.

A recent study by Droin and Kholtei et al demonstrated that the spatial and temporal axes overlap with one another[52]. This is mostly due to the inherently rhythmic behavior of *Wnt* ligand expression. This leads to targets of the *Wnt* signaling pathway such as *Axin2* to exhibit rhythmic patterns of expression. Thus, many established zonated pathways have rhythmic patterns of expression. Among those pathways are many gene sets involved in drug metabolism. Interestingly, while rhythmicity impacted the core pathways that determine zonation, the converse was not true, as the core circadian clock (except for *Cry1*) exhibited no zonation[52]. The overlap with the temporal and spatial axes with the drug metabolism pathways suggests a third axis to consider when describing liver function, chemical perturbation. 2,3,7,8 Tetrachlorodibenzo-*p*-dioxin (TCDD) is a particularly interesting candidate due to its known impact to rhythmicity and zonation pathways[62,66]. It has been established that TCDD in sub-chronic exposures ablates or greatly dampens the oscillations of the majority of core circadian clock genes[66]. This is hypothesized to act through the cis-regulatory action of AhR, TCDD's canonical receptor. AhR is observed to bind at the gene-body of many of the core circadian clock genes two hours post TCDD treatment[66]. Sub-chronic exposure to TCDD also impacts zonation

in a dose-dependent manner[62]. At lower doses, a clear pericentral bias appears across all hepatocytes which reflects a higher activation of pericentral drug metabolism pathways. At higher doses, dysregulation of zonation leads to complete disorganization of most zonation biomarkers[62]. Like in the effects on circadian rhythm, effects on zonation likely stem from the potential cross talk between AhR and *Wnt* signaling pathways[70,71].

While it is established how TCDD may impact both the spatial and temporal organization of the hepatic lobule, it is still less understood how each of these pathways interact with one another. Furthermore, most results discussed above deal with month long exposures to TCDD, and do not describe the initial acute effects that lead to the differential zonation and rhythmicity. Utilizing the model I developed, I show that TCDD, even at acute exposures, has a significant impact on both the zonation and rhythmicity of the liver.

2.2 Results

2.2.1 Visualization of acute hepatocyte response to 2,3,7,8 Tetrachlorodibenzo-*p*-dioxin

To elucidate the transcriptional impact TCDD treatment plays in disrupting hepatocyte rhythmicity and zonation, I used hepatic single-nuclei RNA-seq data from male C57BL/6 mice generated Cholico et al[72]. Mice were housed in a room with a 12:12 light:dark cycle, gavaged with a single dose of 30 μg/kg TCDD (or sesame oil vehicle) at time-point 0 (6:00 AM), after which the livers were collected and snap frozen at timepoints 2, 4, 8, 12, 18, or 24 hours post treatment. The data was then clustered and hepatocytes were identified using established hepatic biomarkers from previous studies[62]. All other cell types were removed from the dataset. After filtering the hepatocytes for low quality and low read count, I was left with 129,373 hepatocytes. The top 15,000 highly variable genes (HVGs) were kept for future analysis. I subsequently performed dimensionality reduction for data visualization and analysis on important TCDD response genes (see section 2.4.1 for more preprocessing details).

I performed a UMAP analysis on the hepatocytes to observe cell clustering patterns (Figure 2.3). I observe that the data naturally clusters between time points (with the exception of 18 and 24 hours with TCDD treatment). Additionally, I find that treatment groups separate sufficiently, indicating significant changes in the gene expression profile with respect to TCDD treatment. Interestingly, the UMAP hepatocyte clusters indicated a circular pattern in expression for both treatments. While UMAP's are known to destroy global distance information[40], I believe this reflects the internal rhythmicity of hepatocytes as this circular UMAP pattern was not found for

other cell types (data not shown). I can also observe while acute TCDD treatment impacts hepatocyte rhythmicity, it does not completely abolish it like in previous studies with sub-chronic TCDD treatments[66].

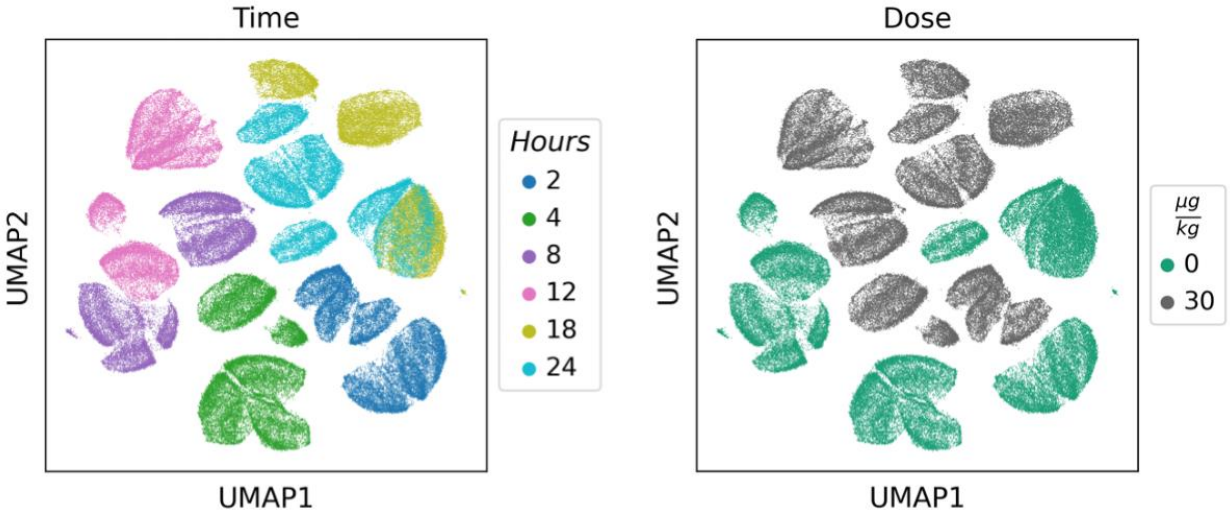

Figure 2.3 Visualization of snRNA-seq for acute TCDD toxicity in mouse hepatocytes. Each dot represents a cell. UMAP of hepatocytes colored by time in hours after treatment on the left, and dose of treatment in μg/kg of TCDD on the right.

To analyze the direct response of TCDD over the time course, I analyze the expression of known TCDD response genes: *Cyp1a1*, *Cyp1a2*, *Ahrr*, and *Tiparp* (Figure 2.4)[73,74]. I find that *Cyp1a1* and *Ahrr* achieve saturation in expression at 12 hours post treatment, while *Cyp1a2* and *Tiparp* achieve saturation in expression sometime at or before two hours post treatment. All response genes exhibited significant increases in expression when treated with TCDD.

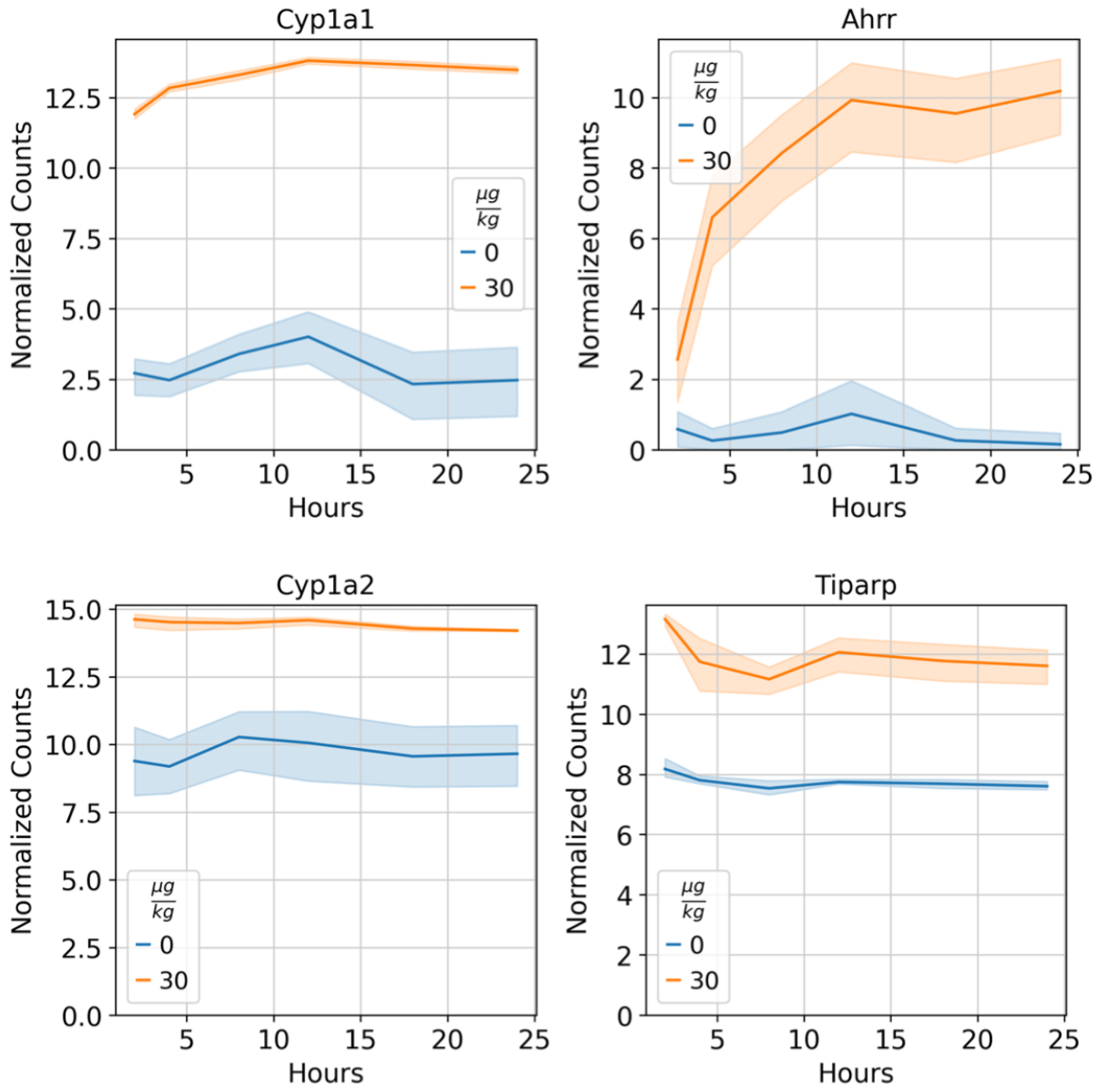

Figure 2.4 Time Series Expression of TCDD Response Genes. Normalized counts (see section 2.4.3) of known TCDD activated genes: *Cyp1a1*, *Cyp1a2*, *Ahrr*, and *Tiparp*. Expression of genes are plotted as function of the hours after treatment by TCDD (30 $\frac{\mu g}{kg}$) or sesame oil vehicle (0 $\frac{\mu g}{kg}$). Shaded regions represent 95% confidence interval.

To analyze the potential acute effects TCDD has on the spatial and temporal metabolism of the liver lobule, I first inferred the porto-central axis for the liver hepatocytes, and then used a mixed non-linear effects model[75] (MNLEM) to classify genes based on what factor (Rhythmicity, Zonation, TCDD treatment) or combination of factors controlled gene expression.

2.2.2 Inferring zonation of hepatocytes from single cell gene expression profiles

Zonation of the liver lobule is not directly measured by snRNA-seq. Thus, zonation needs to be inferred from hepatocyte gene expression. To infer the zonation profile, expression must first be corrected for other confounding factors in the experimental design. For example, known centrally zonated gene *Cyp1a2*[10] is activated by TCDD. Another centrally zonated example, *Slc1a2*, is known to have rhythmic patterns in expression[52]. To correct these confounders and calculate the zonation trajectory I elected to utilize the approach taken by Aizarani et al[76] and Nault et al[62]. Here I first perform batch correction on the single cell data to remove variance stemming from TCDD treatment and time-dependent effects. Then I perform trajectory inference to calculate the latent zonation value of each hepatocyte.

I utilized single **c**ell **V**ariational **I**nference (scVI)[29], a variational autoencoder (VAE), to perform batch correction on the snRNA-seq data (Figure 2.5). The major reason I utilize a variational autoencoder is due to the size of the dataset. After preprocessing, there are ~130,000 cells in the dataset and most single cell integration tools do not scale to datasets of those size easily[77]. scVI can integrate large gene expression atlases and batch correct for variation across many labs[29,77]. Thus, it is expressly tested to scale to millions of cells and can integrate them in a relatively short period of time[29].

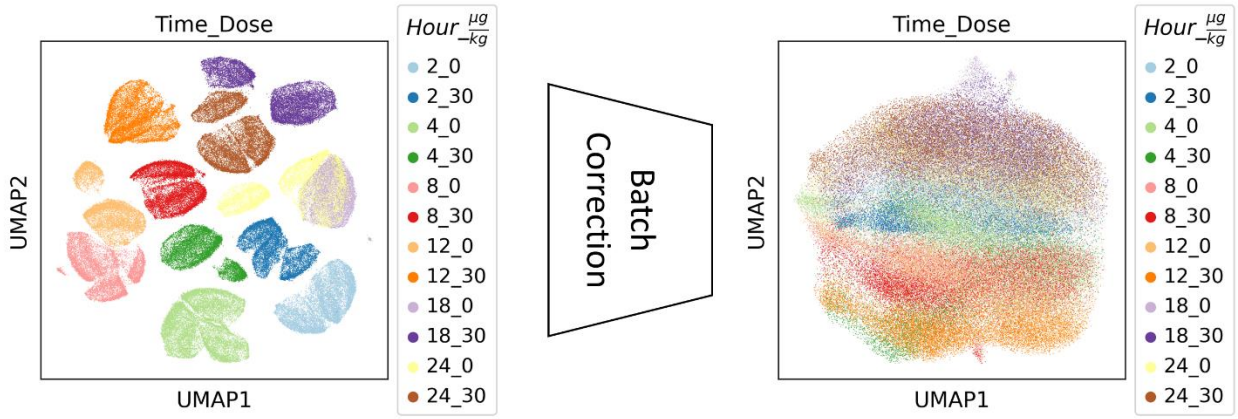

Figure 2.5 UMAP of Batch correction by scVI of snRNA-seq data. UMAP of cells not batch corrected on the left. In the middle is a cartoon of the encoder region of scVI which performs dimensionality reduction. On the right is a UMAP of the latent space of scVI. Cells in both UMAPs are colored by the combined label of hours after treatment and dose of TCDD treatment in μg/kg (Hour_μg/kg).

VAE's like scVI can batch correct for variables using their latent space. When encoding the latent distribution during training, the user can condition the model on some factor in the experimental design (e.g., dose and time). In this way, the model encodes a latent representation that doesn't contain the variance of the factors being conditioned for. I can "denoise" the data by decoding from the latent distribution back into gene expression distribution[29].
Utilizing scVI, I batch correct for TCDD treatment and time of harvest. I then use a trajectory inference algorithm, diffusion pseudo-time[56], on the latent space scVI to infer the trajectory of expression on the porto-central axis. I utilize the second component of the diffusion pseudo-time plot (analogous to components in PCA) as our trajectory (see section 2.4.3) as I observed most zonal genes follow a gradient along this component (Figure 2.6). I will refer to trajectory inference values as pseudo-space, which I define as an ordering of cells based on how closely the cells approximate the expression of central hepatocytes (0 for most portal, and 1 for most central).

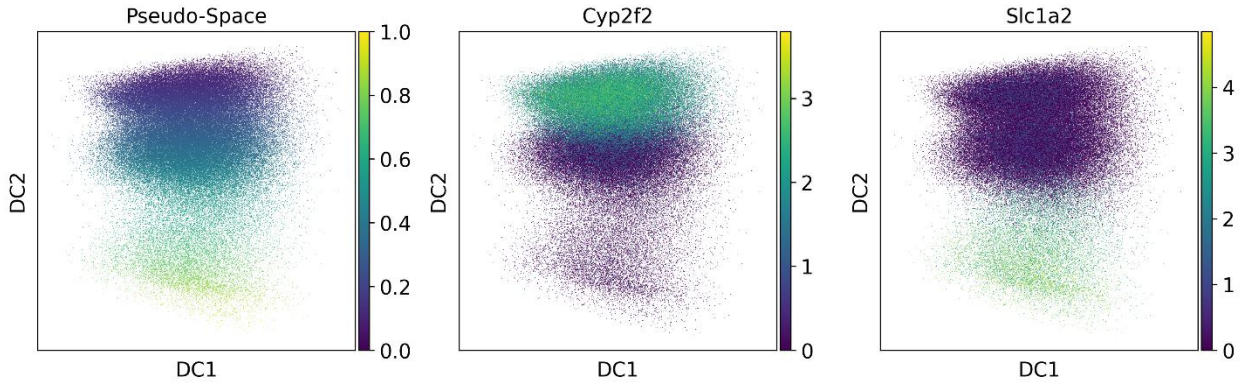

Figure 2.6 Diffusion map visualization of scVI latent space reveals zonation trajectory. Diffusion maps representation of batch corrected latent space of scVI. Each point represents a single hepatocyte. Hepatocytes on the right are colored by pseudo-space metric. Hepatocytes with a pseudo-space value closer to 0 have a more portal hepatocyte-like expression and closer to 1 have a more central hepatocyte-like expression. Hepatocytes in the middle panel are colored by *Cyp2f2* (portal marker) expression. Hepatocytes in right panel are colored by *Slc1a2* (central marker) expression. Expression of each gene is measured in cell normalized counts (see section 2.4.1).

To confirm whether pseudo-space accurately represents the zonation of liver lobules, I plot expression profiles of known zonated genes *Cyp2f2* and *Slc1a2* along the pseudo-space axis. I observe significant correlations between the expression of these genes and the pseudo-space values of the cells (Figure 2.7). As a negative control, I analyze whether I see a similar correlation in known non-zonated genes, such as *Arntl* and *Clock* (Figure 2.7). I show that no significant correlations exist between the pseudo-space metric and these genes.

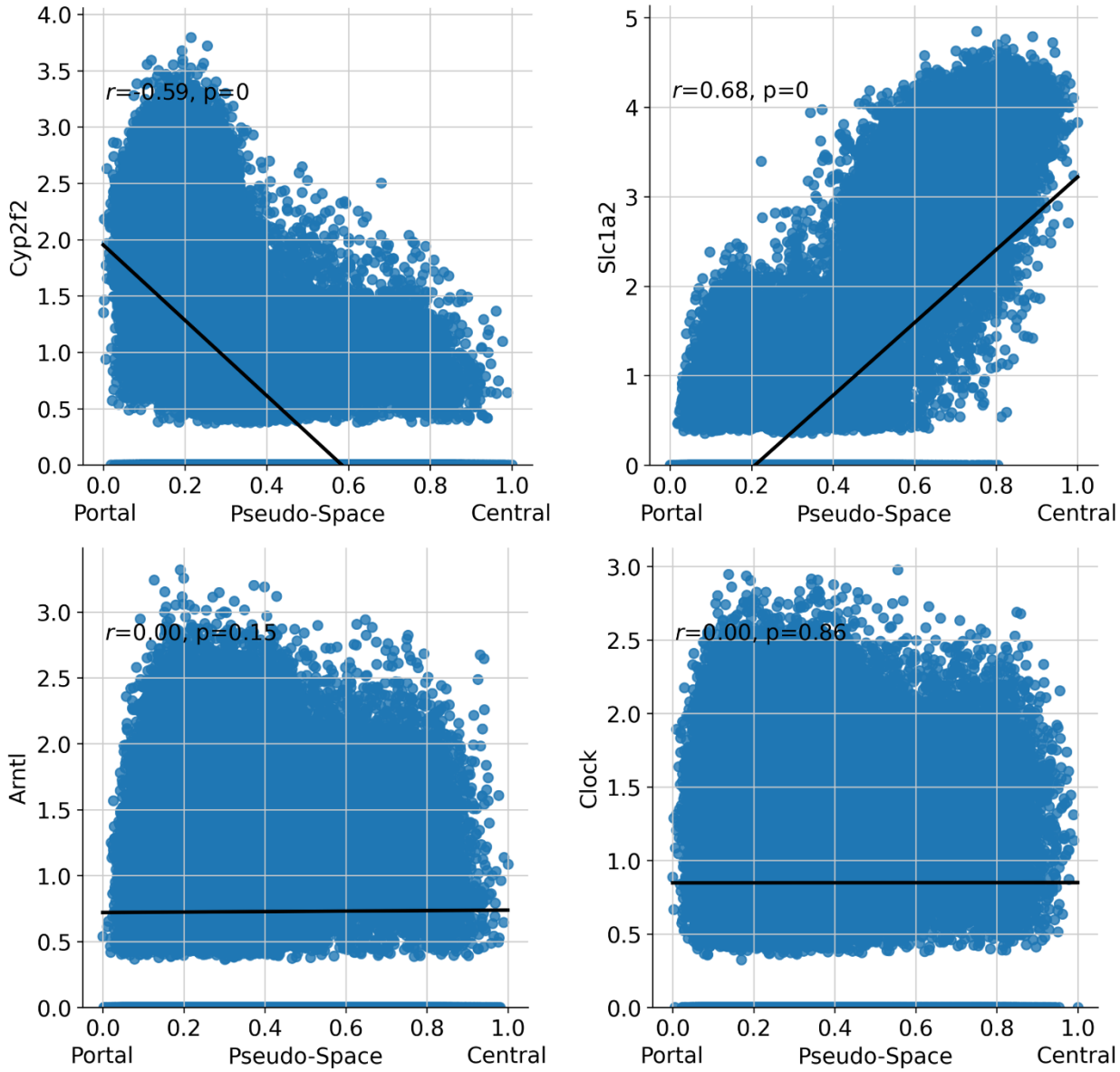

Figure 2.7 Regression plots of cell normalized expression versus pseudo-space. Regression plots of hepatocyte expression *Cyp2f2*, *Slc1a2*, *Arntl*, and *Clock* versus the associated pseudo-space value. Each point is a hepatocyte. Expression of each gene is measured in cell normalized counts. A pseudo-space value of 0 refers to a hepatocyte with more periportal expression. A pseudo-space value of 1 refers to hepatocytes with more pericentral expression. Pearson correlations and their associated p-values were calculated between the gene expression and pseudo-space metric.

To simplify downstream analysis, I bin the pseudo-space values into layers (see section 2.4.3; Figure 2.8). Typically, hepatocytes are separated into three zones in the liver lobule: Zone 1 or

periportal, Zone 2 or mid-lobular, and Zone 3 pericentral cells[53]. However, in more recent studies involving single cell RNA-seq and smFISH show zonation to be finer grained[6]. Selection of the number of layers to bin the cells into is down to the resolution of the data. For example, in Halpern et al[6] there were fifteen zones, while in Droin and Kholtei et al[52], there are only eight zones. I elect to bin into five layers given to make sure there are at least two-thousand cells in each layer. Each bin has an equal length along the pseudo-space axis (e.g., all cells in layer one have a pseudo-space value between zero and one-fifth). Unsurprisingly, I find that there are fewer cells binned into the last (more central) layer than compared to the first (more portal) layer (Figure 2.9). This makes sense since by having smaller radius in the liver lobule, there are fewer cells nearer to the single central vein than nearer to the many portal triads.

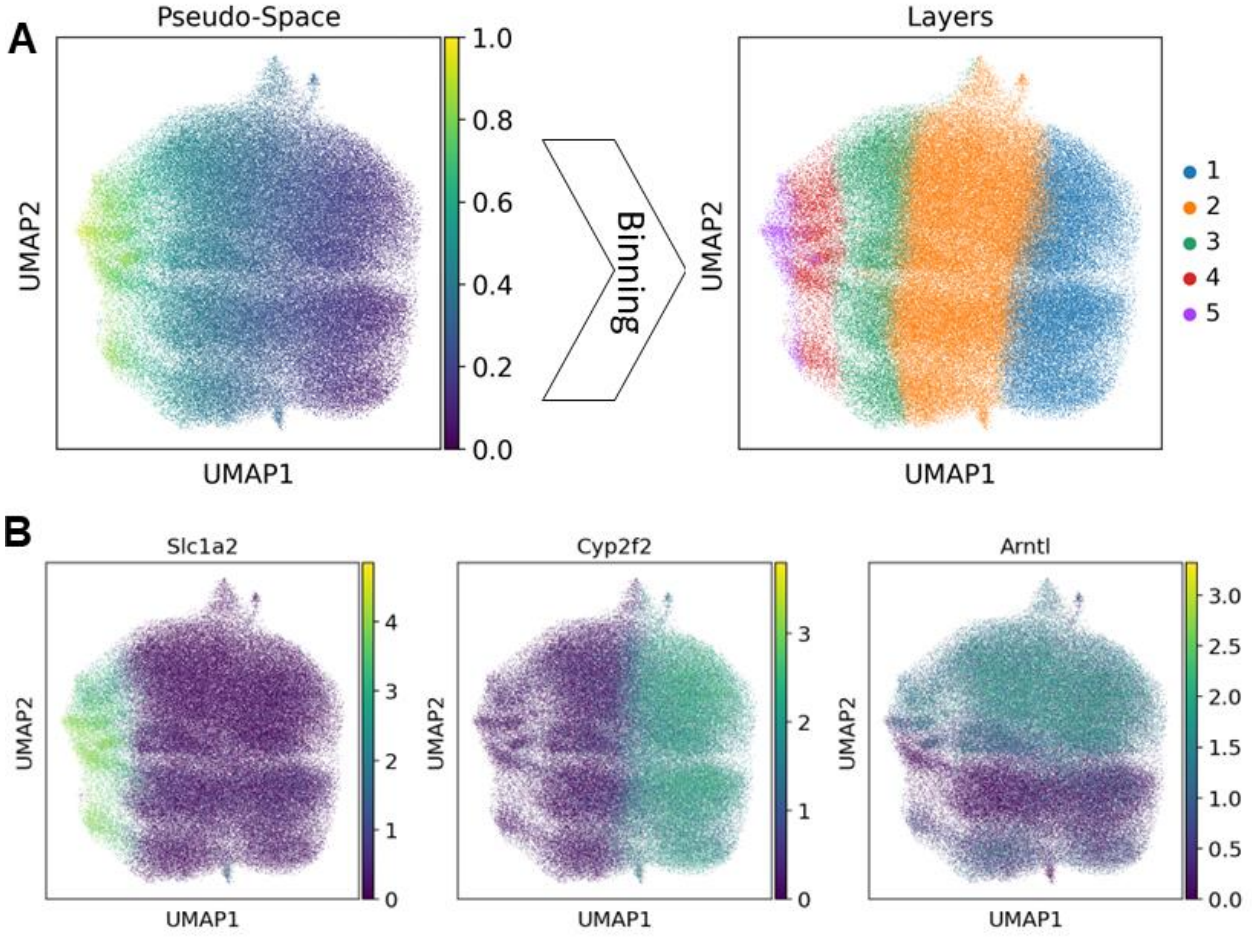

Figure 2.8 UMAP visualization of pseudo-space binning and zonation biomarkers. All UMAPs are of scVI's latent space. A) Top panels represent schematic of binning process. On the top left cells are colored by the pseudo-space metric. A pseudo-space value of 1 refers to hepatocytes with more pericentral expression and a pseudo-space value of 0 refers to hepatocytes

Figure 2.8 (cont'd)

with more periportal expression. On the top right, the same UMAP colored by the binned layers. Each layer refers to hepatocytes with pseudo-space values between (layer number - 1)/5 and layer number/5. B) On the bottom panels are UMAP plots with cells colored by Cyp2f2, Slc1a2, and Arntl gene expression in cell normalized counts.

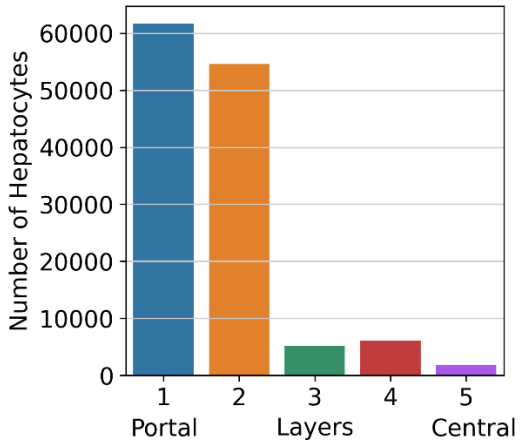

Figure 2.9 Bar plot of the number of hepatocytes in each inferred layer. Layers are described in figure 2.8. Layers with smaller values represent hepatocytes that have expression more similar to pericentral hepatocytes and layers with larger values represent hepatocytes that have expression more similar to periportal hepatocytes.

Using the binned pseudo-space values, I create pseudobulk profiles (sum of counts across cells) for each treatment x time x layer combination (see section 2.4.3). These count sums are then normalized using size-factor estimation to normalize for the number of cells used to create each pseudobulk profile[78]. I use the normalized expression for each treatment x time x layer combination for down-stream analysis and classification of genes.

2.2.3 Classifying hepatocyte genes using a mixed non-linear effects model

To investigate the spatial and zonated expression profiles of genes, and how those profiles are changed by TCDD treatment, I utilize an approach described by Droin and Kholtei et al[52]. In their approach they model rhythmicity (R) using the sum of sine and cosine functions, and zonation (Z) using the first and second order Legendre polynomials. I extend their model to include the influence of TCDD (Dioxin; D) on liver expression. To model the effect of TCDD, I

opted for a simple indicator function in which the hepatocytes were treated with TCDD (D = 1) or were not treated (D = 0). Using these functions, a series of MNLEMs were constructed that represented different combinations of factors (D, Z, and R) that could influence gene expression based on the original experimental design and the inference of zonation (Figures 2.10 and 2.11). Effects from different factors can either be independent of one another (e.g., Z + R) or dependent on one another (e.g., D x Z). How factors can influence one another to change expression is illustrated in figure 2.11. Using this approach, I classified the top fifteen thousand highly variable genes based on the MNLEM that best described said gene's expression (see section 2.4.4 for more details on implementation).

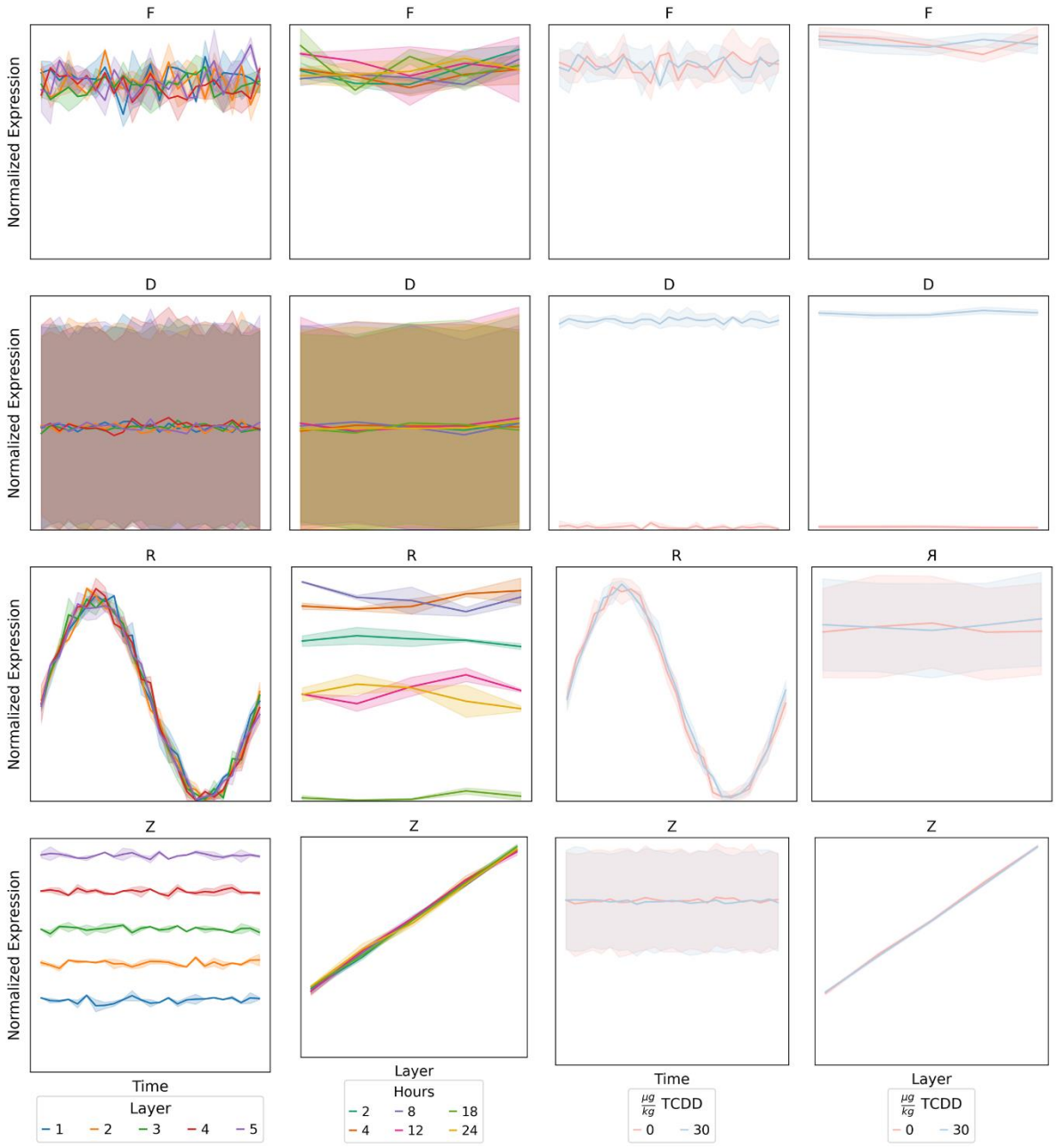

Figure 2.10 Illustrations of each single or no effect models used for classification of hepatocytes genes. Ideal simulations of each model plotted in terms of time (in hours after treatment) versus expression, or inferred layers of the liver lobule. Each line plot is colored either by time, layer of dose of TCDD of treatment. TCDD (Dioxin) influence is delineated by D, influence of liver rhythmicity is delineated by R, and influence of liver zonation (layers) are delineated Z. Models with no influence from either D, R, or Z are delineated as flat, F.

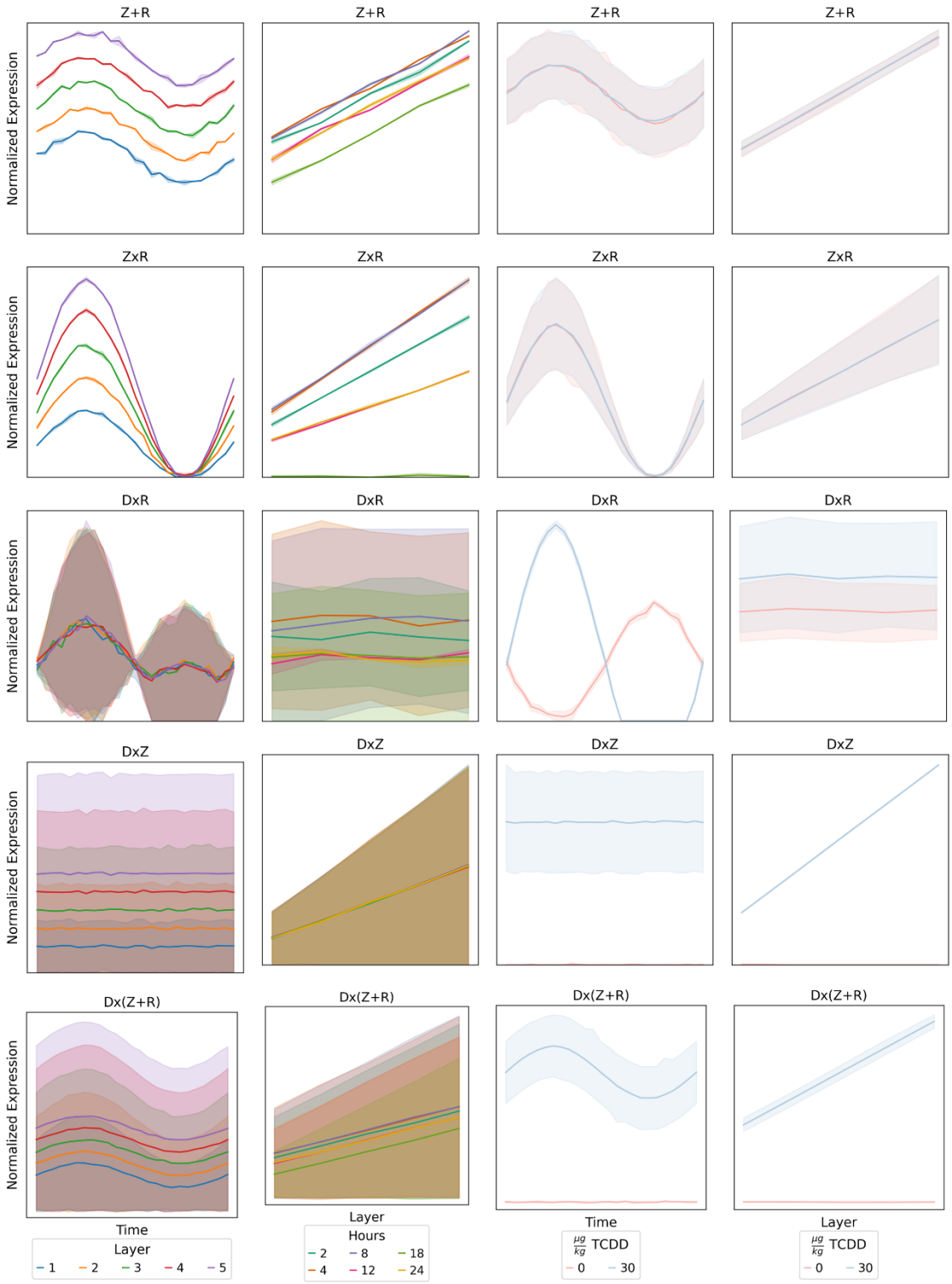

Figure 2.11 Illustrations of multiple effect models used for classification of hepatocyte genes. Ideal simulations of each model as described before in Figure 2.9. Effects that are

Figure 2.11 (cont'd)
separated with "+" indicate models in which factors are independent of one another. Effects that are separated with a "x" indicate models in which factors are dependent on one another.

I first confirm that my classification works by analyzing genes known to have rhythmic, zonal, or TCDD influenced expression. For this I classified Arntl (Rhythmic Gene), and Slc1a2 (Zonated and Rhythmic Gene), and Ahrr (a TCDD response gene that saturates at 12 hours). In figure 2.12 I show that my model can detect the rhythmicity of Arntl and Slc1a2. Additionally, the model can distinguish between TCDD saturation and true rhythmicity. The model is also able to find that Slc1a2 is zonated. Finally, Ahrr is classified as having TCDD influence. I find in this small survey that gene expression profiles reflect the classification of the gene one is looking at (Figure 2.12).

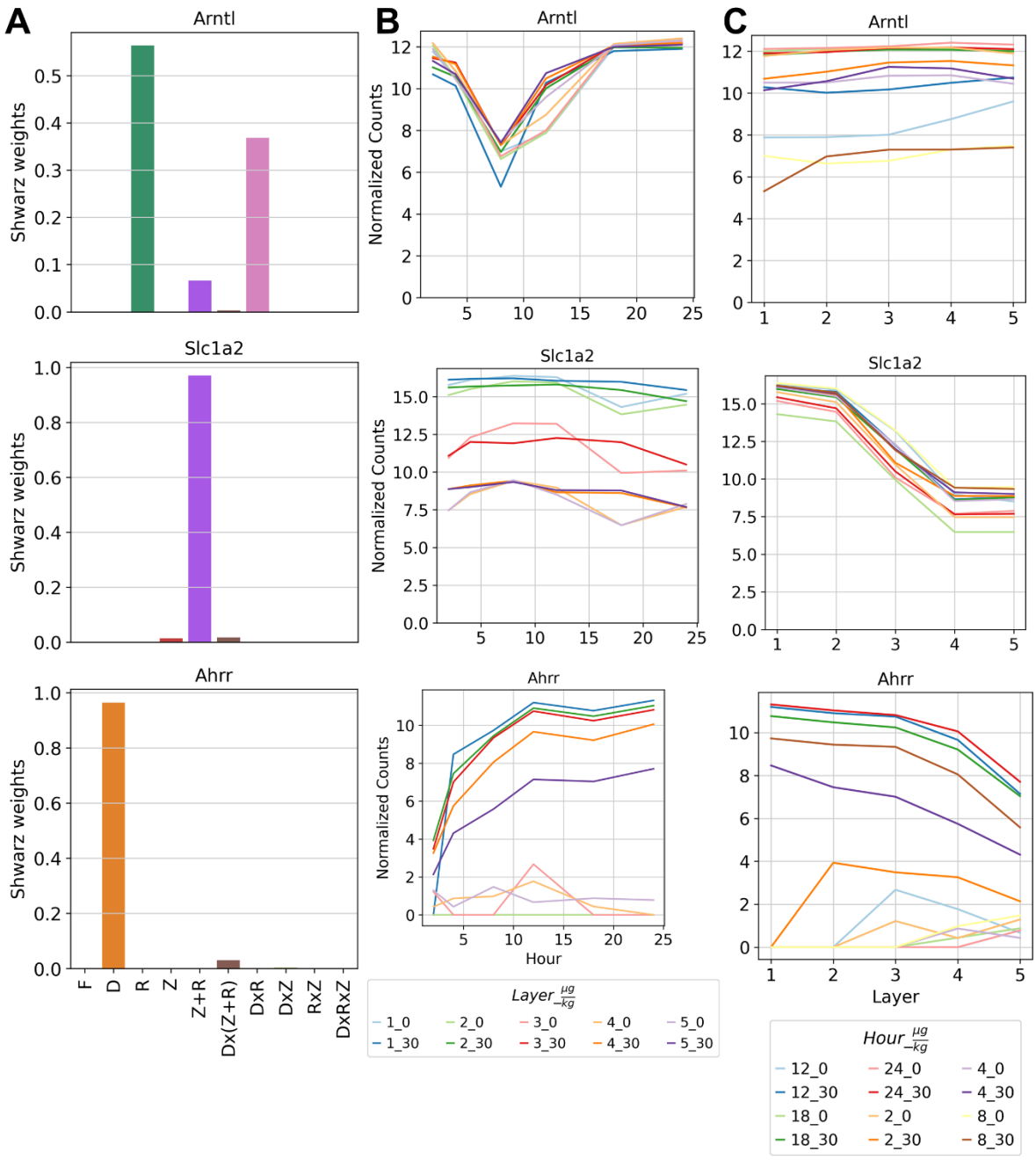

Figure 2.12 Classification of canonical rhythmic, zonal, and TCDD responsive genes. Analysis of canonical rhythmic gene, Arntl, canonical rhythmic and zonal gene, Slc1a2, canonical TCDD responsive gene Ahrr. A) Bar plots of Shwarz weights for each gene classification. The higher the Shwarz weight, the higher the probability the gene belongs in a particular class. Effects that are separated with "+" indicate classes in which factors are independent of one another. Effects that are separated with a "x" indicate classes in which factors

Figure 2.12 (cont'd)

are dependent on one another. TCDD (Dioxin) influence is delineated by D, influence of liver rhythmicity is delineated by R, and influence of liver zonation (layers) are delineated Z. Models with no influence from either D, R, or Z are delineated as flat, F.B) Line plots of gene expression measured in normalized counts (see section 2.4.3). Each line is colored by layer and dose of TCDD (μg/kg), or C) colored by hour after treatment and dose of TCDD.

Looking at the distribution, I find that most genes are not influenced by any factors (Figure 2.13 A). Class sizes for classes that don't include the effect of TCDD agree with previous studies with more genes being zonated than rhythmic[52]. The largest multi-effect class with TCDD influence is Dx(Z+R) (Figure 2.11 A). When I look at the classification of TCDD canonical receptor, AhR, I find that it has a classification of Z+R (Figure 2.13 B). Furthermore, when I perform gene set enrichment analysis (GSEA) on Dx(Z+R) class, I find that it is enriched for many of the canonical pathways involved in TCDD liver response (Figure 2.13 C). I found that no genes were categorized to have dependence between all three factors (DxZxR).

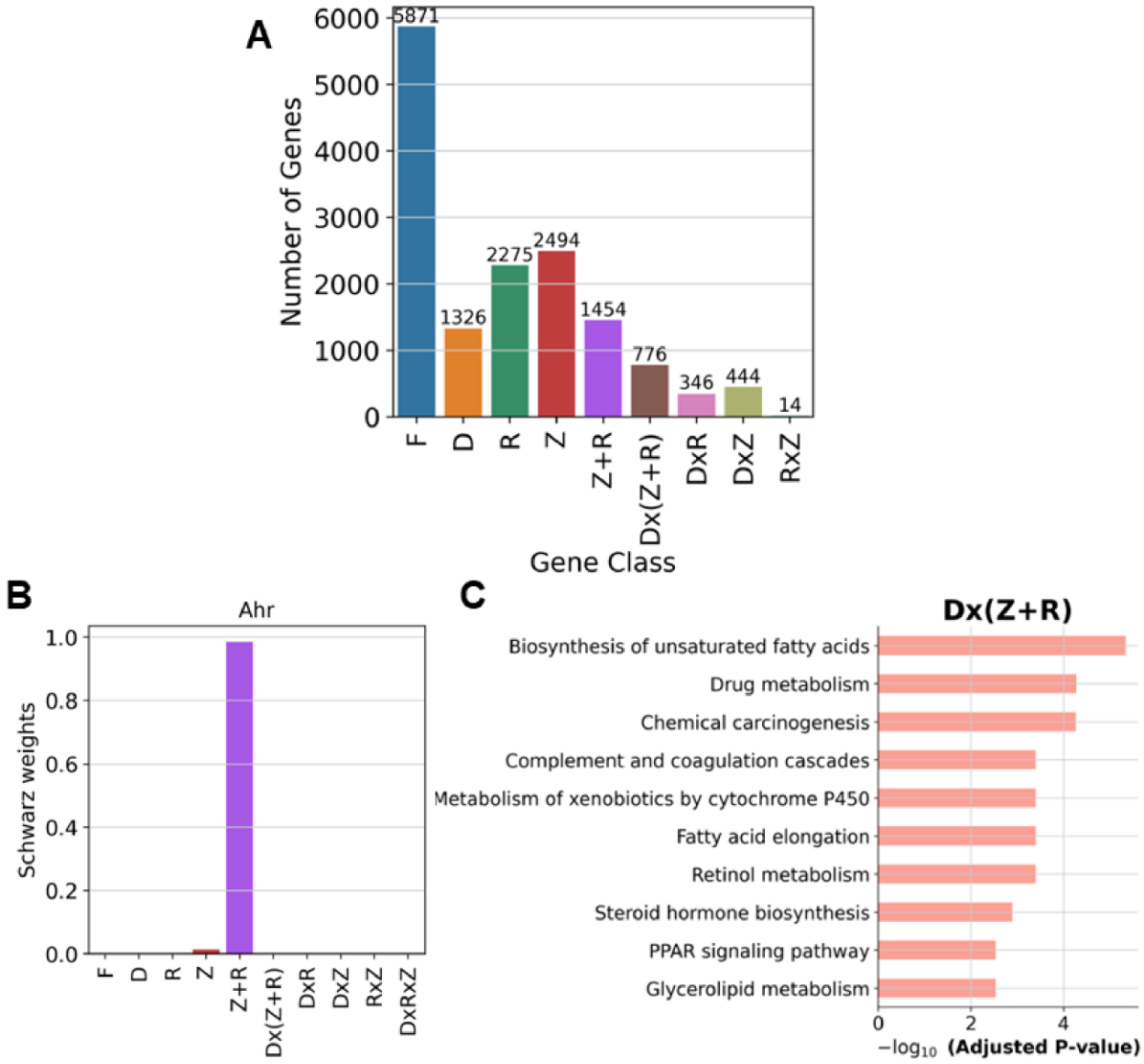

Figure 2.13 Classification distributions and enrichment of TCDD toxicity pathways. A) Bar plot of the number of genes in each class for the top 15,000 HVGs. Heights represent the number of genes in each class. B) Bar plot of Shwarz weights for Ahr classification. The higher the Shwarz weight, the higher the likelihood Ahr belongs to a particular class. C) GSEA analysis of the Dx(Z+R) class of genes. Significance of a particular pathway on the y-axis is represented by its negative log adjusted p-value. The higher the value the more significant the pathway. The top 10 most significant pathways are graphed. I use the above classification scheme to then investigate acute toxicity to rhythmic genes and zonal genes in the liver lobule.

2.2.4 Acute 2,3,7,8 Tetrachlorodibenzo-*p*-dioxin perturbation of rhythmic genes

I first analyzed the impact of TCDD on a core set of circadian rhythm genes (*Clock*, *Arntl*, *Per2*, *Cry1*, *Nr1d1*, *Npas2*, *Rorc)*. To do this I looked at gene classifications for each of the core circadian clock gene. All core circadian genes analyzed contained rhythmicity in their classification. The only genes classified as purely rhythmic were *Arntl*, *Clock* and *Rorc* (Figure 2.14 A). When I looked closely at their expression, I found that *Arntl* and *Clock* exhibit significantly increased expression at 2 and 12 hours (p-value < 0.01 Mann-Whitney U-test) with TCDD treatment. This correlates with the saturation of TCDD response genes analyzed in section 2.2.1 and Figure 2.2 (Figure 2.14 B, C). *Npas2*, *Per2*, and *Nr1d2* all were classified to have TCDD influence (Figure 2.15 A). I find that *Per2* has significantly higher expression for all time points except 18 hours post treatment. *Npas2* has significantly lower expression at timepoint 2, 4 and 24. Finally, *Nr1d2* exhibits significantly higher expression at time point 2 and 4, but significantly lower at time point 12. *Cry1* was classified as zonal and rhythmic, which is corroborated by previous studies[52]. However, I note that the second highest classification of *Cry1* implies TCDD influence which can be seen at time points 18 and 24 (Figure 2.15 A, B).

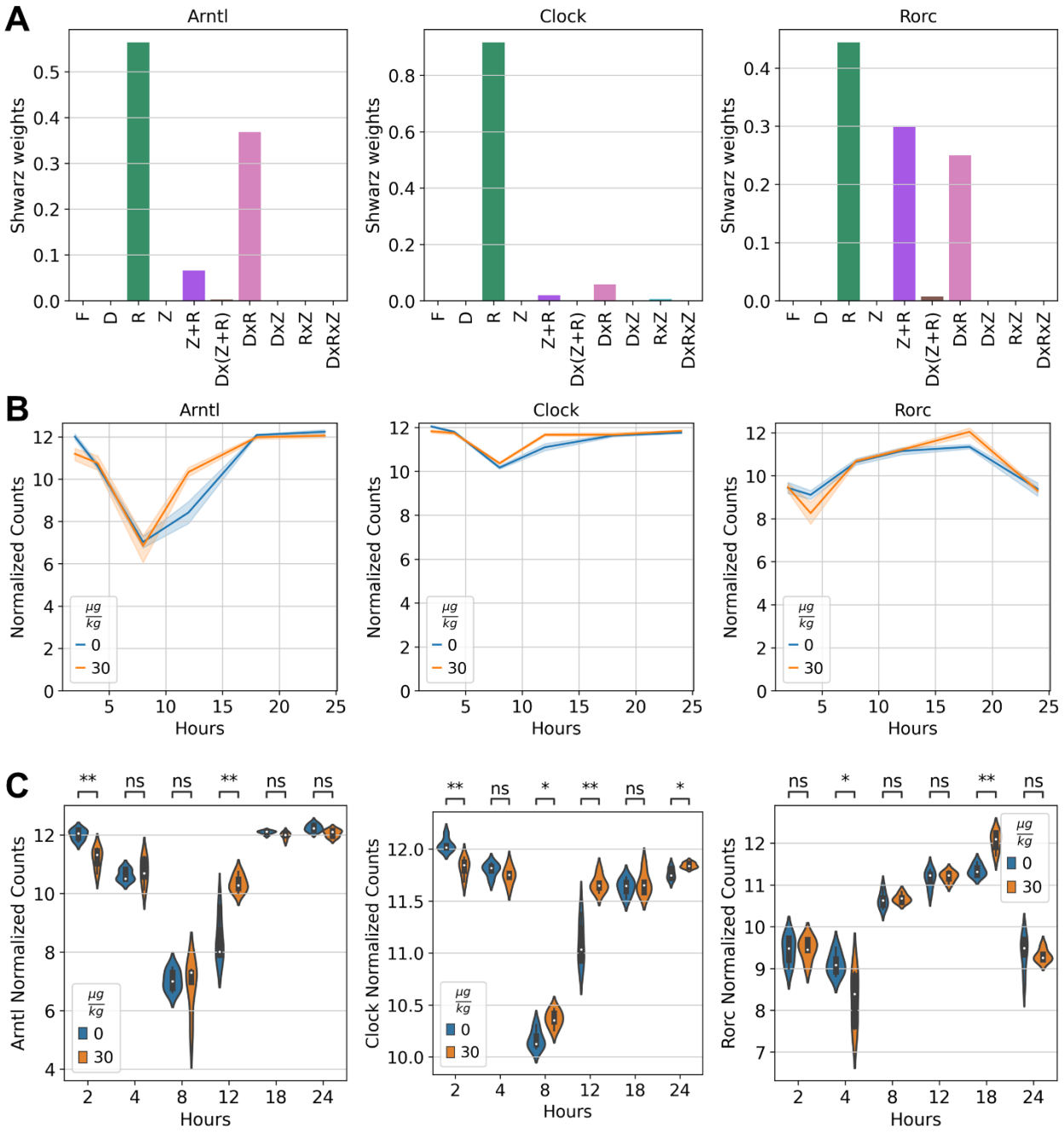

Figure 2.14 Core circadian clock genes classified as only rhythmic. A) Bar plots of Shwarz weights for gene classification of Arntl, Clock, and Rorc. B) Line plots of gene expression for Arntl, Clock, and Rorc graphed with respect to hours after treatment and colored by dose of treatment of TCDD. C) Violin plots plotting the distribution of expression for each treatment condition at each timepoint, with p-value differences described with *'s. ** is 0.001 < p-value < 0.01 and * 0.01 < p-value < 0.05.

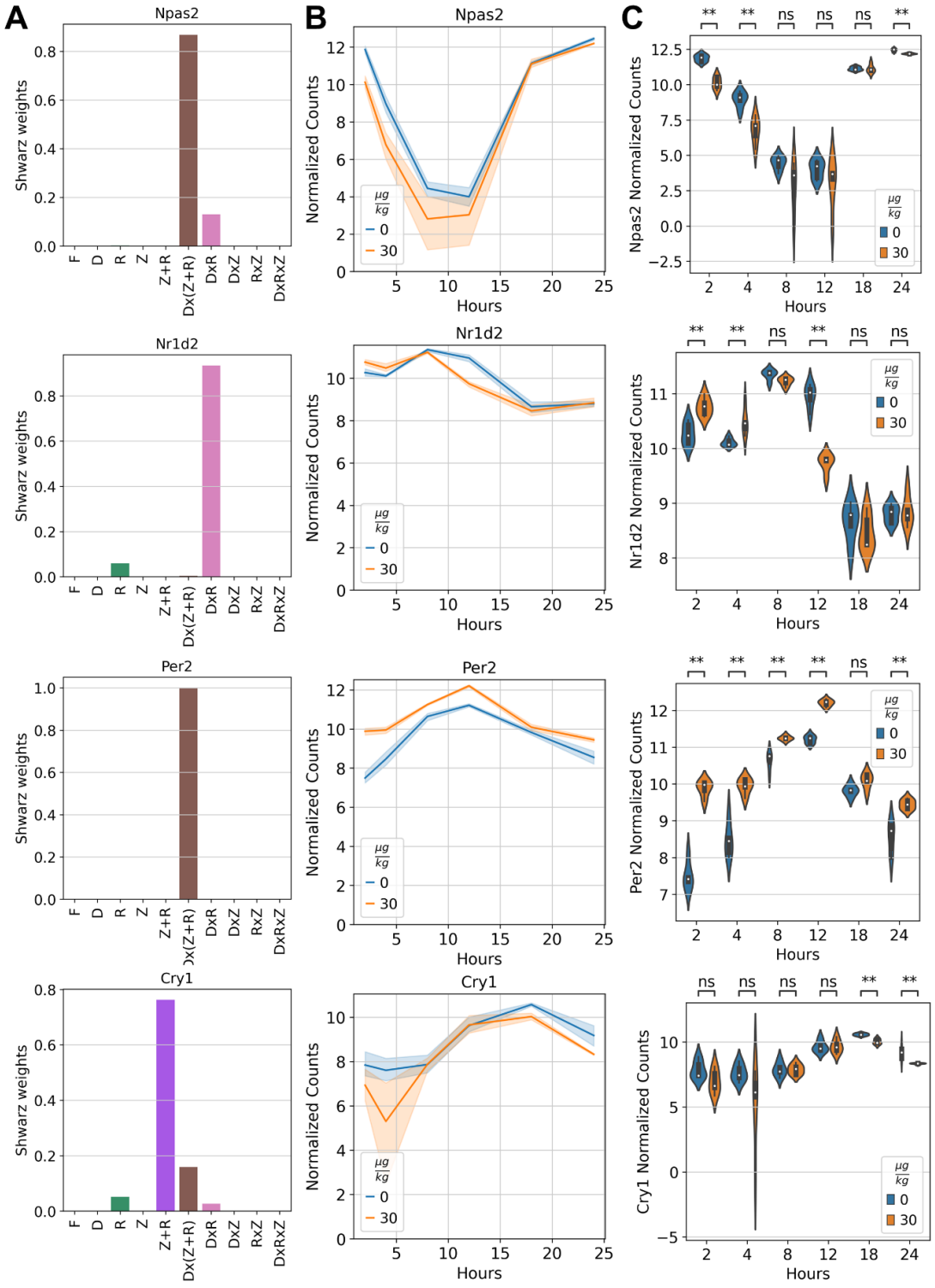

Figure 2.15 Core circadian clock genes classified as having multiple effects. A) Bar plots of Shwarz weights for gene classification for Npas2, Nr1d2, Per2, and Cry1. B) Line plots of gene

Figure 2.15 (cont'd)

expression for Npas2, Nr1d1, Per2, and Cry1 graphed with respect to hours after treatment and colored by dose of treatment of TCDD. C) Violin plots of Npas2, Nr1d1, Per2, and Cry1 plotting the distribution of expression for each treatment condition at each timepoint, with p-value differences described with *'s. ** is 0.001 < p-value < 0.01 and * 0.01 < p-value < 0.05.

I next looked at the impact of TCDD on rhythmic genes in DxR and Dx(R+Z) classes by first trying to see if TCDD has induced or removed rhythmicity from a gene (see section 2.4.5). To investigate whether a gene has gained or lost rhythmicity with TCDD treatment, I use the likelihood ratio test (see section 2.4.6). Here I compare whether the expression of a gene better fits a rhythmic or non-rhythmic (flat) model. Genes that fit rhythmic model in control but fit flat model in treatment are classified to have lost rhythmicity. Genes that fit a flat expression model in control but fit a rhythmic expression model in treatment are classified to have gained rhythmicity. If the gene fits a rhythmic model in both treated and control conditions, they are classified to have kept rhythmicity. Looking at the distribution of genes in the TCDD influenced rhythmic classes, I see that 13% of genes have gained rhythmicity, while 21% of genes have lost rhythmicity (Figure 2.16 A). When I perform GSEA on genes that lost rhythmicity, I find hallmarks of TCDD gene expression response such as "Metabolism by CYP450" and "Chemical Carcinogenesis" (Figure 2.16 B)[79]. When I perform GSEA on genes that have gained rhythmicity, I find another hallmark of TCDD expression response, retinol metabolism (Figure 2.16 B)[80].

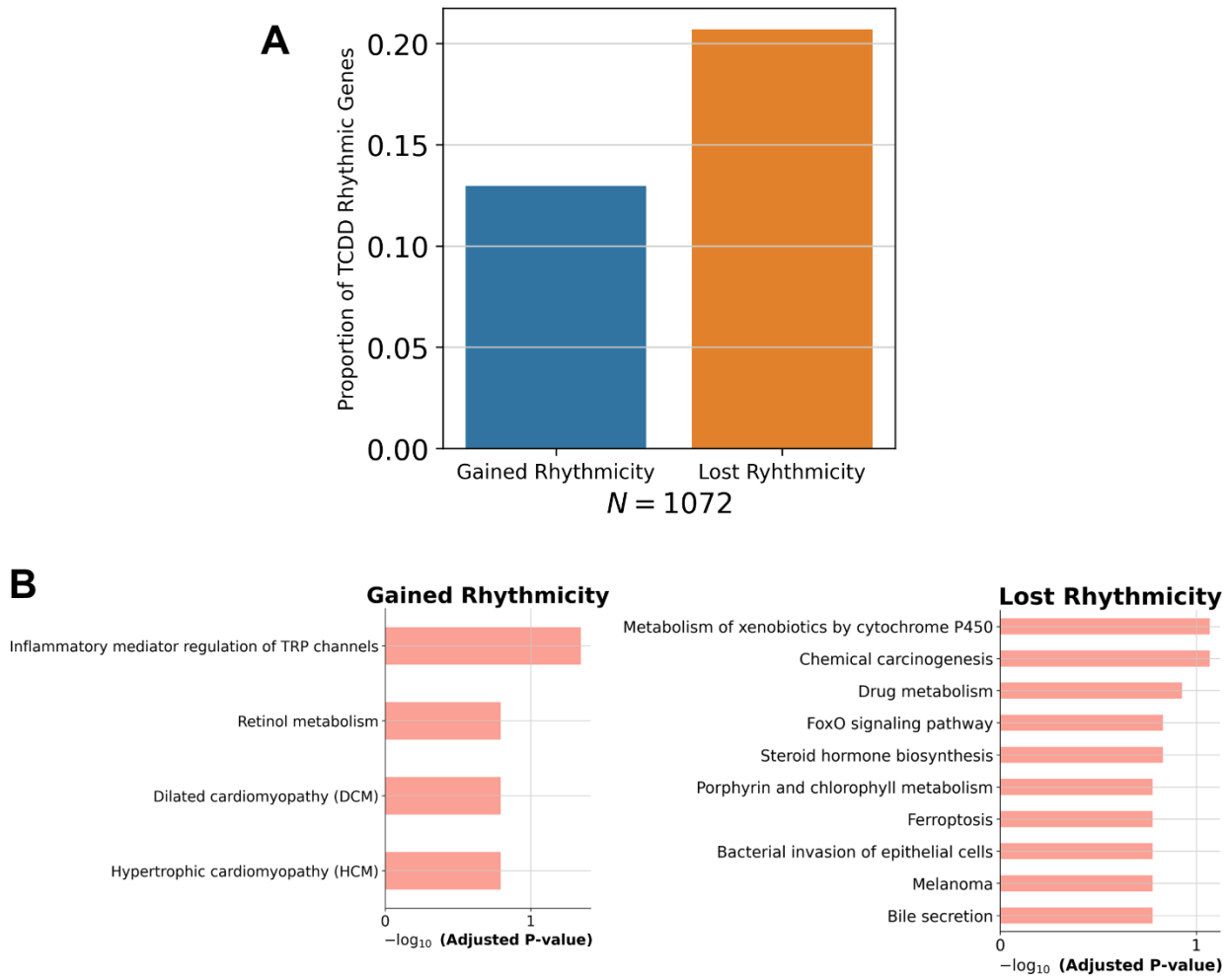

Figure 2.16 Analysis of gain/loss of rhythmicity for TCDD influenced rhythmic genes. A) A total of 1072 genes that were classified to be influenced by rhythm and TCDD were analyzed as to whether they had gained, or lost rhythmicity based on the likelihood ratio test. A bar plot of the proportion of genes analyzed colored by whether they lost (in orange) or gained rhythmicity (in blue). B) GSEA analysis of gene sets for genes that gained and lost rhythmicity. Significance of a particular pathway on the y-axis is represented by its negative log adjusted p-value. The higher the value the more significant the pathway. The top 10 most significant pathways are graphed unless fewer than 10 pathways were significant.

Looking at the remaining genes that kept rhythmicity, I did analysis on how TCDD impacted the properties of their rhythm. In this case I analyzed the amplitudes and phases of the genes and how they changed with TCDD treatment. When analyzing the core circadian clock genes, I only see small changes in the phase and amplitude of genes (Figure 2.17 B). When looking across all

genes that kept rhythmicity, I similarly see that there are no major trends reflecting a delay in phase or reduction of amplitude (Figure 2.17 C). When looking at whether the magnitude of the genetic change with respect to TCDD treatment had a correlation with magnitude of the change of phase or amplitude to TCDD treatment, I found a weak correlation (Figure 2.15 D).

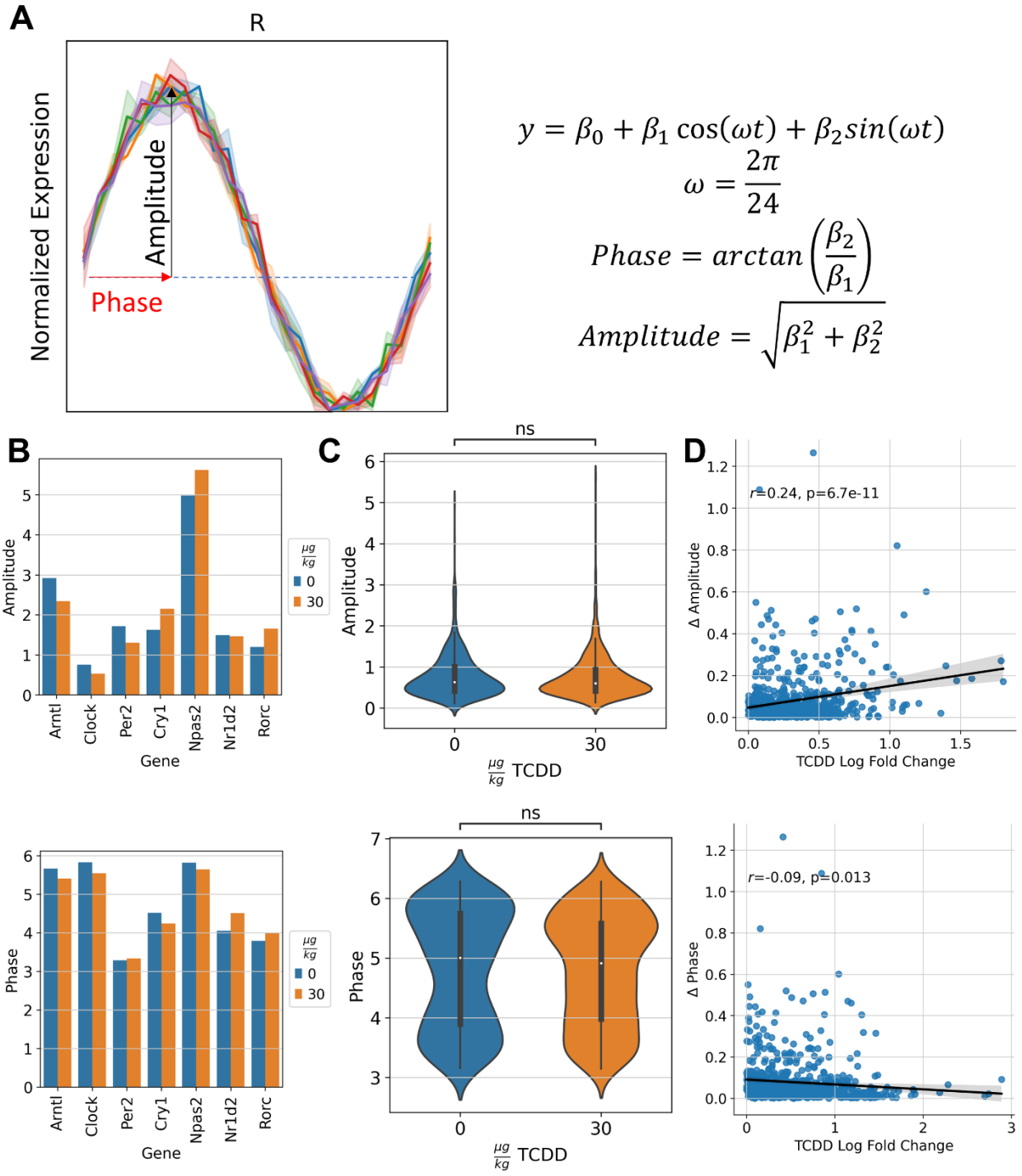

Figure 2.17 Analysis of rhythmic parameters for TCDD influenced rhythmic genes. A) Schematic and equations describing phase and amplitude of gene's expression. B) Bar plot of core circadian clock genes (Arntl, Clock, Per2, Cry1, Npas2, Nr1d2, and Rorc) and their respective phase and amplitude in each treatment group. Phase is measured in radians, and

Figure 2.17 (cont'd)

amplitude is measured in normalized counts. C) Violin plot of all genes that kept their rhythmicity's rhythmic parameters. D) Regression plot of the magnitude of change of phase or amplitude vs. the magnitude of the mean log fold change in expression with respect to treatment. Each point represents a single gene.

2.2.5 Acute 2,3,7,8 Tetrachlorodibenzo-p-dioxin perturbation of zonal genes

Zonation is primarily determined by the Wnt/β-catenin pathway[81] and influenced by the *Ras* pathway and hypoxia pathways[6]. To investigate if TCDD influences these pathways in a meaningful way, I have taken a list of known targets of each pathway[6,52,82] to see if they are enriched in TCDD influenced gene classes. I find that all zonation pathway targets are enriched in the Z+R class of genes. For Z+R's TCDD influenced counterpart, Dx(Z+R), and the dual effect DxZ class, I find that all pathway targets except targets for down regulated by the *Ras* pathway were enriched (Figure 2.18). From this I conclude that the core pathways of zonation are impacted by acute TCDD treatment.

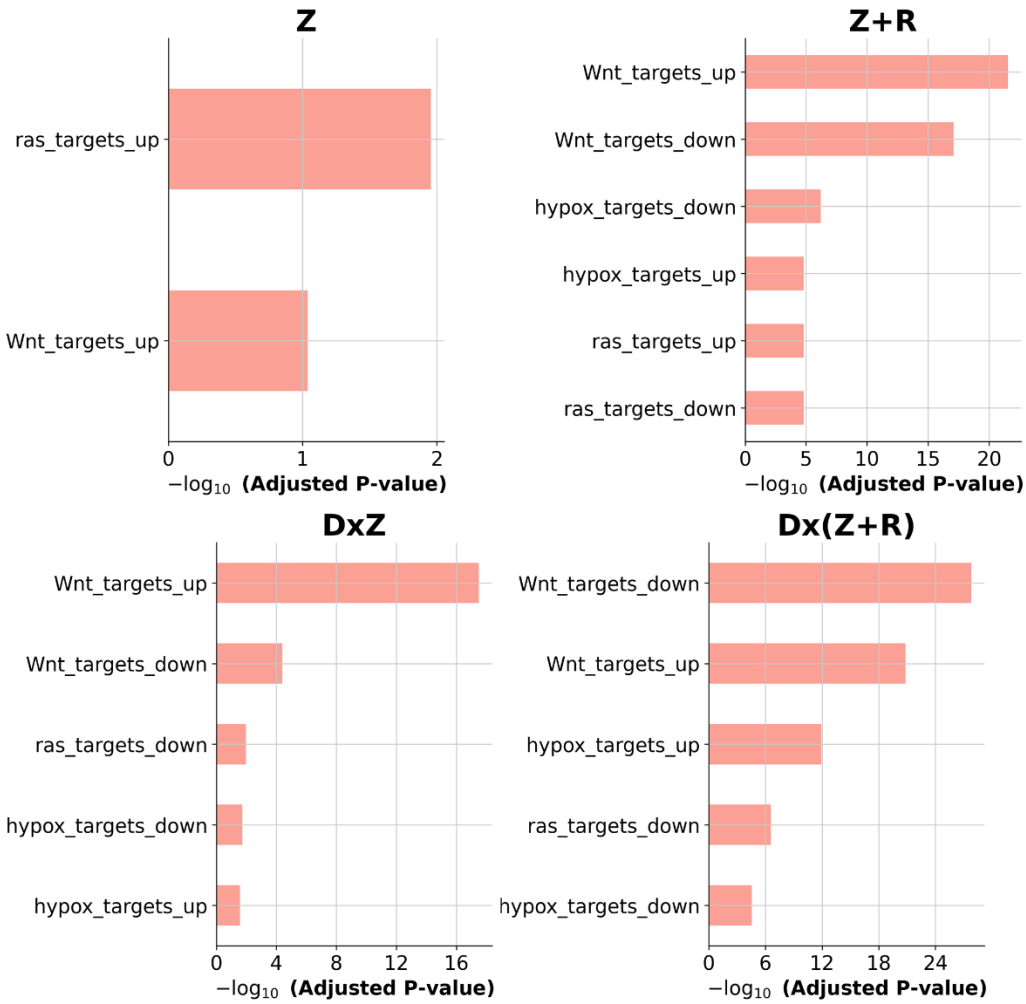

Figure 2.18 Enrichment analysis of zonation regulation pathways on all zonation gene classes. GSEA of pathways that regulate zonation (Wnt, Ras, and hypoxia pathways)6,67. Both Z (Zonation) and Z + R (Zonation + Rhythmicity) represent zonated genes not affected by D (Dioxin). DxZ and Dx(Z+R) represent zonated genes affected by D. Significance of a particular pathway on the y-axis is represented by its negative log adjusted p-value. The higher the value the more significant the pathway.

Similar to their rhythmic counterparts, I analyzed whether genes in the DxZ and Dx(Z+R) have either gained or lost zonation. Like with the rhythmic likelihood ratio test, I performed a zonated analogue (see section 2.4.5). I find that 18% of genes analyzed had gained zonation when treated with TCDD, and 13% of genes analyzed had lost their zonation with TCDD treatment (Figure 2.19 A). When I performed GSEA on each of the groups, I found that UDP-glucuronosyltransferase related pathways such as “Pentose and Glucoronate Interconversions”

were enriched in genes that lost zonation (Figure 2.19 B). I found no enriched pathways for the genes that gained zonation. Finally, when analyzing whether the genes that lost or gained zonation were more portally enriched or centrally enriched, I found that genes that lost zonation were more centrally enriched even when compared to background (p-value < 0.001 Kolmogorov-Smirnov Test) (Figure 2.19 C).

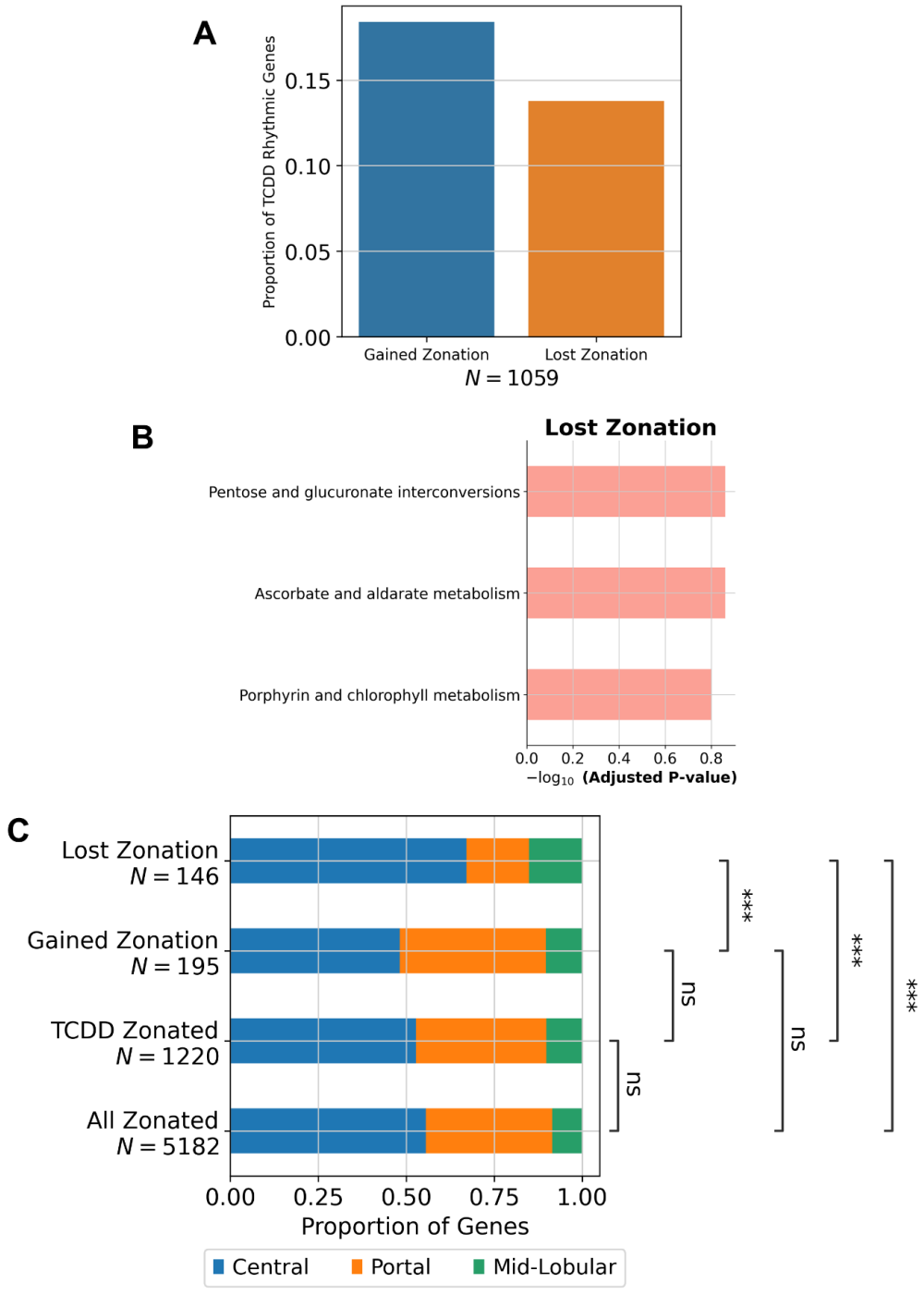

Figure 2.19 Analysis of loss or gain of zonation based on TCDD treatment. A) Bar plot of proportion of 1059 TCDD influenced zonated genes. B) Enrichr[83] analysis of genes that lost zonation with TCDD treatment. C) Stacked bar chart describing the distribution of zonated genes and in which zone of the liver lobule are those genes most highly expressed. Zonal location

Figure 2.19 (cont'd)

based on the maximum expression of gene. Differences in distribution calculated using the two sample Kolmogorov-Smirnov test.

Finally, looking at the genes that kept zonation, I looked at how TCDD impacted zonation parameters. For this I calculated the line of best fit for layer dependent expression of each gene. I call the slope of the line of best fit the gene's "zonation slope". I filtered out genes with centers of expression in the mid-lobular region, as they have a non-monotonic zonation pattern and thus are not linearly zonated. I found that linearly zonated genes that had larger magnitude log fold changes with respect to TCDD treatment also had larger magnitude changes in their zonation slope (Figure 2.20).

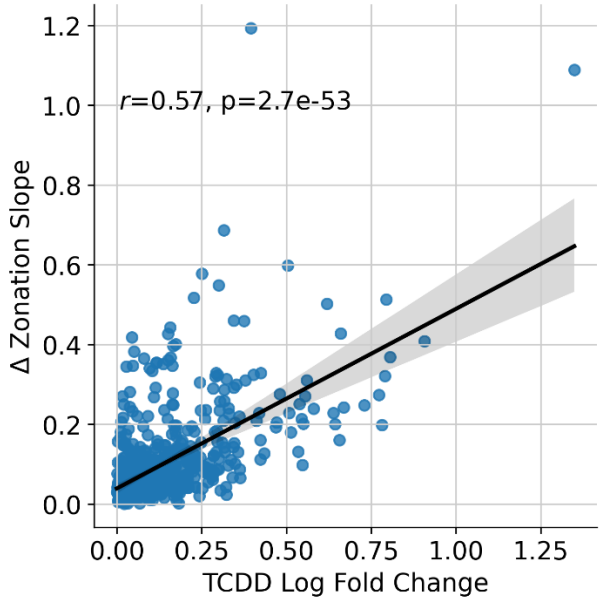

Figure 2.20 Regression plot of TCDD induced log fold change vs. TCDD induced change in zonation. Each dot represents a zonated gene as determined by classification. Magnitude of change of the slope of the line best fit across liver layers is plotted versus the magnitude of the mean log fold change in expression of the gene.

2.3 Discussion

Analysis of drugs in the liver have often been limited by having to study the chronobiology and spatial zonation separately. In this chapter, I have taken previous results integrating chronobiology and spatial zonation in the liver and have extended them to account for chemical dependent influence. This remains important as how the liver reorganizes its metabolic functions

in response to toxic stimuli remains understudied. I show that even in acute situations, TCDD has a significant impact on both the rhythm and zonation of the liver. I show that AhR, TCDD's canonical receptor, has zonal and rhythmic components to its expression, and that many effects of TCDD also have zonal and rhythmic components.

Previous studies showed that sub-chronic treatment by TCDD at the same dose caused a large dampening effect on the amplitude, and a large shift in the core circadian clock[66]. While this isn't true for acute perturbation, I still observe significant TCDD influence in nearly half of the core circadian clock genes I analyzed. This corroborates previous results that AhR directly impacts the core-circadian feedback loops[66]. Most genes in the core-circadian clock analyzed that were classified as impacted by TCDD were downstream CLOCK-ARNTL (e.g., *Per2*, and *Nr1d2* binding (with the single exception of *Npas2*). It has been shown in previous studies that AHR, ARNTL, and CLOCK colocalize on *Per2* and *Nr1d1*, suggesting that AhR may interrupt normal CLOCK-ARNTL binding. Additionally, βNF-activated AhR has been shown to interact with ARNTL in Hepa1c1c7 cells impairing CLOCK-ARNTL heterodimer formation at E-boxes within the Per1 promoter[84]. *Per1* was filtered out of the dataset due to low variance, however I see signs of *Per1* inhibition in *Per1*'s cell normalized expression from time points 4-12 hours as well as 24 hours post TCDD treatment (Figure 2.21).

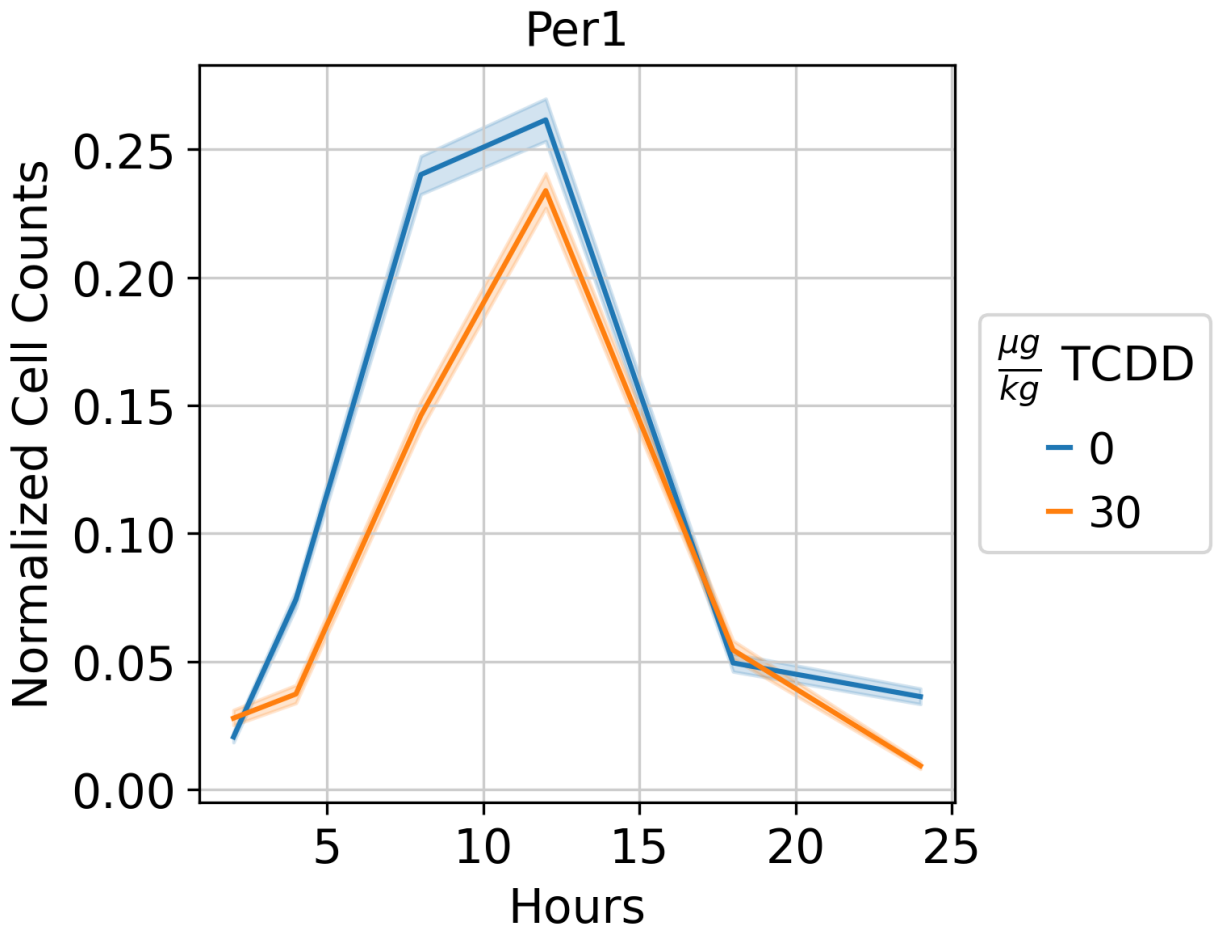

Figure 2.21 Time series expression of Per1. Cell normalized counts of Per1 plotted with respect to hours post oral gavage at ZT0 (6:00 AM). In blue (0 $\frac{\mu g}{kg}$ TCDD) are measurements from hepatocytes treated with just sesame oil vehicle. In orange are measurements from mice treated with 30 $\frac{\mu g}{kg}$ TCDD. Areas in shaded region represent a 95% confidence interval.

TCDD's overall effect on all rhythmic genes in the hepatocyte genome remains mixed. Most genes had modest effects on their expression at specific time-points. Additionally, the overall shape of the oscillations for genes that kept rhythmicity with treatment remained relatively constant. However, genes that gain or lose rhythmicity make up around 15% of all rhythmic genes. Additionally, GSEA results of the genes that lost rhythmicity had many of the hallmarks of TCDD gene expression response[79,85]. This suggests that TCDD ablates rhythmicity in many of the major pathways it targets.

TCDD also significantly impacts zonation pathways in the liver. TCDD alters the zonation of nearly a quarter of all zonated genes. Of those genes, around thirty percent of them have either gained or lost zonation. Genes that have lost zonation are enriched in phase II UDP-glucuronosyltransferase metabolism. Additionally, genes that have completely lost their zonation are enriched in the central liver of the liver lobule. This correlates with the central enrichment of AhR in the liver lobule[59].

Interestingly, I show that genes significantly affected by TCDD are enriched for targets of regulators of zonation pathways. This is potentially due to the cross talk between the *Wnt*/β-catenin signaling pathway[70]. Interestingly, much of this enrichment is in Dx(Z+R) class. The rhythmic component of this class has been hypothesized to be linked to *Wnt* signaling by non-parenchymal cells in the pericentral zone[52]. Potentially, TCDD activation of AhR enriched in the central layer[59] could cause perturbation to this signaling, and thus explain the observed perturbation to rhythmicity in these genes. Further analysis of the spatial and rhythmic effects of TCDD on non-parenchymal cells would be needed to confirm this hypothesis.

Unanswered by my analysis is the exact role AhR plays in the perturbation of the spatial and rhythmic axes of the liver lobule. Are the effects I see on rhythmicity, for example, a result of TCDD activating AhR and AhR binding directly to cis-regulatory elements, or is this a more hierarchical effect? Is TCDD inducing changes to the core circadian clock, which in turn is inducing changes in the rhythmicity of other genes? Is a downstream effect of TCDD's perturbation of rhythmicity the perturbation of zonation as many zonation regulators are also rhythmic? Or are the effects completely independent of one another? Is it different for different gene pathways? One way to probe these questions in a high throughput manner is through cis-regulatory binding analysis of AhR and its potential binding partners (ARNTL and ARNT) through technologies like ChIP-seq. The presence or absence of binding for each of these pathways could help delineate which pathways are directly perturbed by canonical AhR and which pathways are being perturbed by non-canonical TCDD toxicity pathways.

The liver is a fantastically complex organ with a distinct space-time logic to its metabolic organization. Since the liver is central to the metabolism of many drugs, understanding how drugs impact this space-time logic may help to 1) describe how hepatocytes reorganize in response to chemical perturbation, and 2) how break down of this logic is related to liver injury. Here I present a methodology for analyzing the effects of chemical perturbations on zonation and

rhythmicity. The methodology can be extended to dose-response experiments via a change in the function describing the effect of chemical treatment. Additionally, the methodology can easily be applied to any snRNA-seq data set involving chemical perturbation of the liver. I envision pharmacologists and toxicologists utilizing the methodology described here to describe not only the impact that chemical have on particular pathways, but also how those impacted pathways figure into the overall metabolic compartmentalization of the liver.

2.4 Methods

2.4.1 Preprocessing of single-nuclei RNA-seq preprocessing

Cholico et al[72] performed clustering and cell type annotation of the dataset as described in previous lab studies[79,86]. All preprocessing was performed using the *scanpy.pp* pacakage[87]. Raw counts were normalized to the median total cell count using the *normalize_total*. They were then log transformed with a pseudocount using the *log1p* function. I refer to counts normalized in this way as "normalized cell counts". Highly variable genes were selected using the *highly_variable_genes* function.

Filtering of cells first started with the removal of all non-parenchymal cell types so that I was only left with hepatocytes. Hepatocytes with fewer than 1,500 counts and 3 genes were removed from the dataset. Cells with fewer than 200 genes being expressed were also removed. Highly variable genes were selected using the *highly_variable_genes* function.

2.4.2 Batch correction using scVI

scVI[29] was trained on the snRNA-seq data to perform batch correction on the data. To remove sample wise batch variance, I labeled all the cells in each biological sample and then used that label as input for the batch labels in scVI. I used default parameters were used for the structure and training the model which are shown in table 2.1 and 2.2 respectively.

Hyperparameter	*Value*
Latent dimension	30
Number of layers	1
Layer width	128
Dropout rate	0.1
Kullback-Leibler weight	$5 * 10^{-5}$
Gene expression distribution	NB
Latent distribution	Normal

Table 2.1 Hyperparameters for scVI's variational autoencoder model.

Hyperparameter	*Value*
Training epochs	46
Learning rate	0.001
Learning rate decay	10^{-6}
Optimizer	Adam
Optimizer epsilon	0.01

Table 2.2 Hyperparameters for scVI's variational autoencoder training.

2.4.3 Layer calculations

The latent space representation of the cell normalized counts was used as input into the Diffusion maps algorithm. Diffusion maps was calculated using the function *diffmap* from the *scanpy.tl* python package[56]. The second component of the diffusion maps representation was taken and then min-max scaled in order to generate the pseudo-space metric. The metric was oriented so that centrally enriched expression had the highest values and portally enriched expression had the lowest values.

Cells were binned into five layers using the pseudo-space metric. Each bin represents a fifth of the pseudo-space trajectory (i.e., cells in bin i have pseudo-dose values between $\frac{i-1}{5}$ and $\frac{i}{5}$).

Counts from each treatment x time x layer combination were summed across all cells in then

normalized using the *computeSumFactors* function from the *scran R* package[78]. I call these pseudo-bulk normalized counts "Normalized Counts" in the manuscript.

Genes were classified based on which layer they had maximum "Normalized Counts". Genes that had maximum expression in layers 1 and 2 were considered portal. Genes that had maximum expression in layer 3 were considered mid-lobular. Genes that had maximum expression in layers 4 and 5 were considered central.

2.4.4 Design of Mixed Non-Linear Effects Models

Mixed non-linear effects models (MNLEM) were deployed using the *MixedLM* class in *statmodels* python package[88]. In Table 2.3 I describe the individual terms for each factor: TCDD influence (D), rhythmicity (R), and zonation (Z). Equations for each class described in terms in Table 2.3 are described in Table 2.4.

Term	***Effect***	***Equation***
D	D	$\begin{cases} 0\ if\ Sesame\ Oil\ Control \\ 1\ if\ TCDD\ treatment \end{cases}$
R_{Sin}	R	$\sin(\omega t)$
R_{Cos}	R	$\cos(\omega t)$
Z_{P1}	Z	l
Z_{P2}	Z	$\frac{3l^2}{2}$

Table 2.3 Terms for mixed linear effects models. Each term is denoted by its name and its effect. D is TCDD (Dioxin) Influence, R is rhythmicity, and Z is zonation. t is the time in hours after treatment. l is the layer the of the liver lobule. ω is the conversion factor between t and radians which is equal to $\frac{\pi}{12}$.

Class	*Equation for Model*
F	$y = \beta_0$
D	$y = \beta_0 + \beta_1 D$
R	$y = \beta_0 + \beta_1 R_{Sin} + \beta_2 R_{Cos}$
Z	$y = \beta_0 + \beta_1 Z_{P1} + \beta_2 Z_{P2}$
Z+R	$y = \beta_0 + \beta_1 R_{Sin} + \beta_2 R_{Cos} + \beta_3 Z_{P1} + \beta_4 Z_{P2}$
RxZ	$y = \beta_0 + (\beta_1 R_{Sin} + \beta_2 R_{Cos}) * (\beta_3 Z_{P1} + \beta_4 Z_{P2})$
DxR	$y = \beta_0 + \beta_1 D * (\beta_2 R_{Sin} + \beta_3 R_{Cos})$
DxZ	$y = \beta_0 + \beta_1 D * (\beta_2 Z_{P1} + \beta_3 Z_{P2})$
Dx(Z+R)	$y = \beta_0 + \beta_1 D * (\beta_2 R_{Sin} + \beta_3 R_{Cos} + \beta_4 Z_{P1} + \beta_4 Z_{P2})$
DxZxR	$y = \beta_0 + \beta_1 D * (\beta_2 R_{Sin} + \beta_3 R_{Cos}) * (\beta_4 Z_{P1} + \beta_5 Z_{P2})$

Table 2.4 Equations for each mixed non-linear effects model used for classification. Each term is denoted by its name and its effect. D is TCDD (Dioxin) Influence, R is rhythmicity, and Z is zonation. y represents the predicted gene expression. β represents the associated gene weight. All other terms such as R_{Sin} and Z_{P1} are defined in Table 2.3.

Implementation of the MNLEM was almost identical to Droin and Kholtei et al[52]. These equations were fit to normalized count of each individual gene using the Nelder-Mead optimization algorithm[89]. A noise offset ($\sigma_0 = 0.15$) was added to the data to make sure that overfitting was avoided. Equations with the smallest overall Bayesian information criterion[90] (BIC) were classified with their corresponding class. BIC acts as a general multi-comparison analogue to the likelihood ratio (χ^2) test as it penalizes more complex models. The exception to classifying with models that have the smallest BIC were when models tied with one another. I defined a tie as having a relative difference of 1%. In the case of ties, models with fewer parameters were selected.

Bar graphs of classification were calculated using Shwarz weights. Shwarz weights are calculated using differences between the BIC values and the minimum BIC value in across all models (BIC_{Min}) using the following equation:

$$shwarz\ weight = \frac{\exp \frac{BIC_i - BIC_{Min}}{2}}{\Sigma_{i=0}^{n}(BIC_i - BIC_{Min})}$$

2.4.5 Differential Rhythmicity and Zonation

To perform differential rhythmicity and zonation analysis, I fit gene expression to models much like what I did in section 2.4.5. However instead of using Nelder-Mead, I used ordinary least squares with the *ols* function from the *statsmodels.api* python package. For each gene, I separated expression values between treated and control. These groups of expression values were then fitted either to the R class model or F class model (see table 2.4 for description of functions) in the case of differential rhythmicity or the Z class model or F class model in differential zonation. To see if the R or Z class models were more descriptive of gene expression, I used the likelihood ratio test (also called the χ^2 test). Models that were gene expression that was significantly better explained by the R model was said to have rhythmicity (the same goes for zonation and the Z model). The likelihood ratio test was implemented using the *chi2* function from the *scipy.stats* python package.

I estimated rhythmicity parameters from the fitted parameters in R models. If I let $a = \beta_1$ and $b = \beta_2$ then I can calculate the amplitude of the expression oscillation as:

$$Amplitude = \sqrt{a^2 + b^2}$$

And I can define the phase as:

$$phase = atan2(b, a)$$

Linear zonation slope was calculated for all zonal genes that kept zonation post TCDD treatment and were not mid-lobular. To calculate zonation slope I fit a simpler zonation model than the one in table 2.4:

$$y = \beta_0 + \beta_1 l$$

I fit this model much in the same way I did above during differential zonation. The zonation slope was equal to β_1. The procedure was similar to how one would find the line of best fit.

2.4.6 Statistical Tests

I determined differential expression between treatments and parameter values using the Mann-Whitney U-test using the function *mannwhitneyu*. To determine if there were differences between the distribution of gene zones between gene lists, I used the Kolmogorov-Smirnov test implemented in the *ks_2samp* function. All functions used for evaluating statistical significance were in the *scipy. stats* python package[91]. Gene set enrichment analysis was performed using *Enrichr*[83] pathway analysis with in the *gseapy* package[92]. All significant pathways had an FDR < 0.2.

CHAPTER 3
PREDICTION OF SINGLE DOSE CHEMICAL PERTURBATIONS ACROSS CELL STATES USING VARIATIONAL AUTOENCODERS

3.1 Introduction

The problem I aim to tackle in this chapter of the thesis is generalizing chemical responses across cell types. More formally, given that I know the response for some chemical X in cell types A and B, can I then use that information to accurately predict the response for the same chemical in a third cell type C? This is an important but challenging task in pharmacology and toxicology. Biological context makes the physiological response to chemical perturbation unique for different cell types. Thus, the information about the responses in cell types A and B may not be directly applicable to the response in cell type C. As a result, toxicologists require a great deal of information across biological systems to make assessments about the safety of a particular drug or chemical. A technology that could help address this problem is single-cell RNA-Seq (scRNA-seq). Perturbational scRNA-seq can measure the chemical responses of tens of thousands of individual cells across thousands of genes[93]. As a result, I can start to map the transcriptomic space of chemical perturbations at the single cell level. However, even with advances like the Mix-Seq protocol[94] or comprehensive datasets like Sciplex[95], scientists have only explored a small portion of the cell type x chemical perturbation space. Given this challenge, computational biologists have attempted to leverage large datasets to predict chemical perturbations in a variety of biological contexts[96]. Specifically, a number of deep generative modeling tools have been developed to predict chemical perturbations across cell types[30,33,41,97,98]. Such models have the potential to quantitatively map relevant portions of the cell type x chemical perturbation space using a relatively small amount of data.

The first use of deep generative modeling in the prediction of chemical perturbations was scGen[30], a variational autoencoder that performed dimensionality reduction by encoding scRNA-seq data into a latent space. It then used linear vector arithmetic to predict the state of virtual cells in the latent space, and finally decoded the virtual cells back to the full gene expression space. Alternative approaches have also been described in the literature. For example, scPreGan[98] is a generative adversarial network which utilizes a generator-discriminator framework to predict the distributions of the unknown perturbed cell type. Another alternative is CellOT[97], which is an autoencoder framework that utilizes optimal transport to predict cell type-specific perturbation. However, I find existing approaches perform poorly when approximating

perturbations in *in vivo* experiments. More specifically, when used to predict the *in vivo* expression of cell-type dependent differentially expressed genes (DEGs), they are unable to approximate the expression means accurately. As such, there is a need for models that can predicts the chemical perturbation of cell type specific responder genes appropriately.

Here I propose **single cell Variational Inference** of **Dose-Response** (scVIDR), a regression-based improvement to the scGen model (Figure 3.1). scVIDR boosts prediction accuracy for cell type specific differentially expressed genes (DEGs). Additionally, scVIDR approximates high-dose experiments better than other state of the art algorithms. The model accomplishes this across several datasets including mouse liver cells treated with 30 μg/kg TCDD sub-chronically (Nault et al)[79,86], PBMCs treated with IFN-β (Kang et al)[99], and different mammalian species treated with LPS6 (Hagai et al)[100].

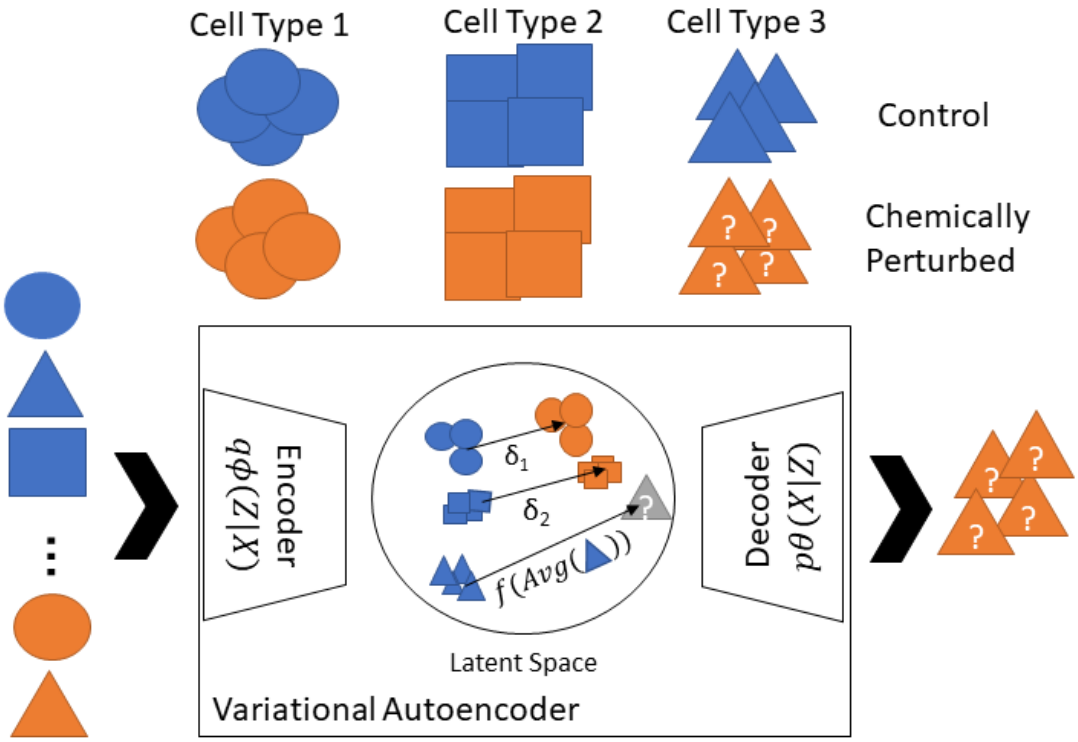

Figure 3.1 Schematic of scVIDR for prediction of a single dose for some unknown cell type. Outline of the scVIDR model for expression prediction for unknown single dose-response in cell type 3. Training is done using cell types 1 and 2 as input to a variational autoencoder model. The difference between the means of latent representations of the control and treated groups, δ_1 and δ_2, are used as input into a linear regression model, which then predicts the unknown δ_3 of cell

Figure 3.1 (cont'd)

type 3. I then use the decoder portion of the model to output the latent space predictions back into gene expression space.

3.2 Results

3.2.1 Description of scGen training and prediction

I begin by considering a scRNA-seq dataset $X = \{x_i\}_{i=1}^{N}$ consisting of N cells, where x_i represents the expression profile of cell i. I 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 I know X and not z, I 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, I use a neural network, the encoder, to approximate it using a different Gaussian distribution, $q_\phi(z|X)$. To map values back from z to X, I 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 (see section 3.4.1 for full mathematical description).

I characterize whether a cell has been treated with a set concentration of the chemical of interest with the indicator variable t (Figure 3.1 A). I set $t = 1$ for cells that have been treated with the chemical (treatment) and $t = 0$ for cells that have not been treated (control). A 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

$$\hat{z}_{i,A,t=1} = z_{i,A,t=0} + \delta$$

where $z_{i,A,t=0}$ is the latent gene expression representations of cell type A,[30] and δ is the difference between the centroids of the treated and control training groups in the latent space. scGen is described as a fixed model[98], meaning that δ is not conditional on the cell type, i.e. it assumes that perturbations in the latent space are consistent across all cell types. While this may hold for datasets with conserved responses[30], in datasets with more heterogenous tissues I find that perturbations in the latent space are not consistent across cell types. Thus, scGen is not

effective for chemical perturbation prediction when evaluating responses across highly distinct cell types.

3.2.2 scGen's δ deviates greatly from cell specific differences when outlier cell types are present.

The scGen model assumes that perturbations on the latent space are similar in magnitude and direction. To define what I mean by this, I first need to define a cell specific perturbation vector δ, δ_c. I define the difference between the latent centroids of the treated (t = 1) and control (t = 0) groups for a particular cell type A as $\delta_{c=A}$ or as:

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

where $\bar{z}$ is the average on the latent space.

I calculated δ_c using a model trained on a snRNA-seq dataset of livers from mice gavaged with 30 μg/kg TCDD sub-chronically (Nault et al)[79]. A UMAP projection of the data can be seen in Figure 3.2, where I can observe by eye that perturbations on individual cell types differ significantly. However, since UMAP only preserves local rather than global distances, I need a more formal way of examining the data.

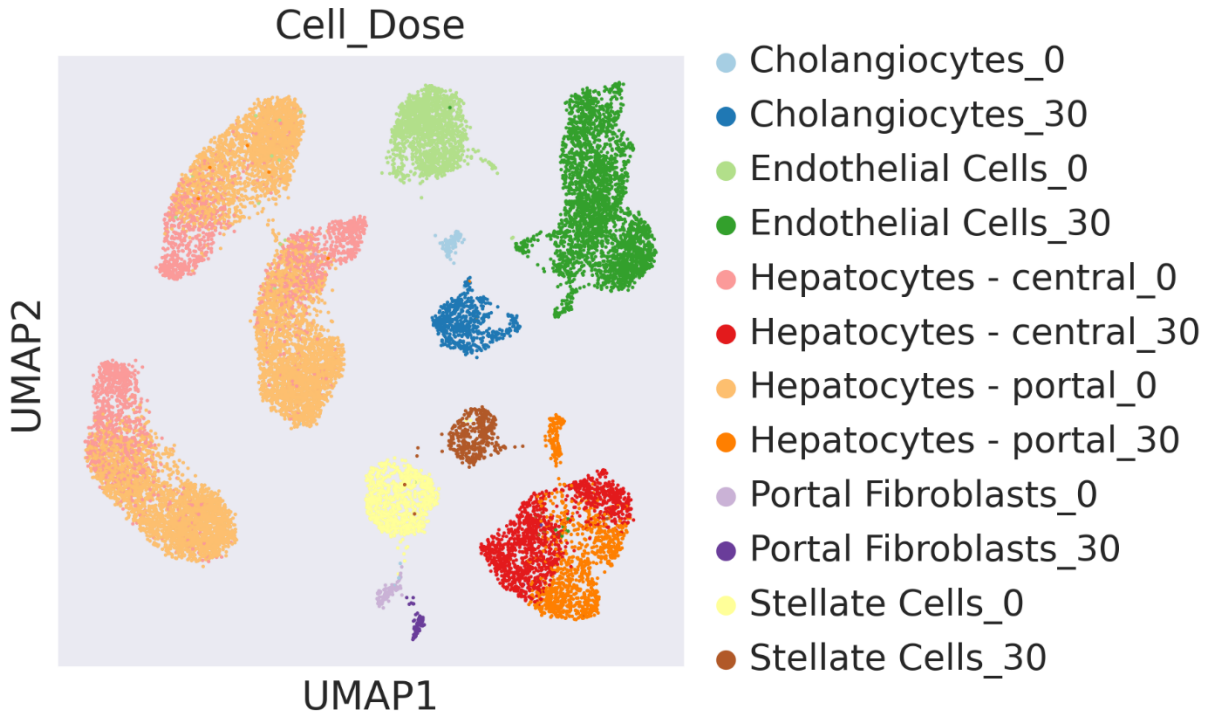

Figure 3.2 UMAP projection of snRNA-seq data from TCDD-perturbed mouse livers. UMAP is performed on normalized gene expression of each cell. Individual points represent

Figure 3.2 (cont'd)

UMAP of gene expression for individual liver cells. Cells are colored by cell type and dose (μg/kg) of TCDD.

I evaluate the assumptions of scGen more formally using a combination of dimensionality reduction that preserves global distances, and high dimensional metrics. In Figure 3.3A, I use PCA to show that δ_c for many of the cell types in the dataset deviate significantly from δ_{scGen}. δ_{scGen} is defined as the difference between the average of treated cell subtracted by the average of the control cells on the latent space of the scGen model. Specifically, δ_{scGen} has a higher overall magnitude than most other cell types except endothelial cells (Figure 3.3 B). To evaluate whether the directions for δ_c are consistent with δ_{scGen} I calculated the cosine distance between δ_{scGen} and δ_c, which showed that δ_{scGen} deviates significantly from all δ_c in terms of direction δ_{scGen}.

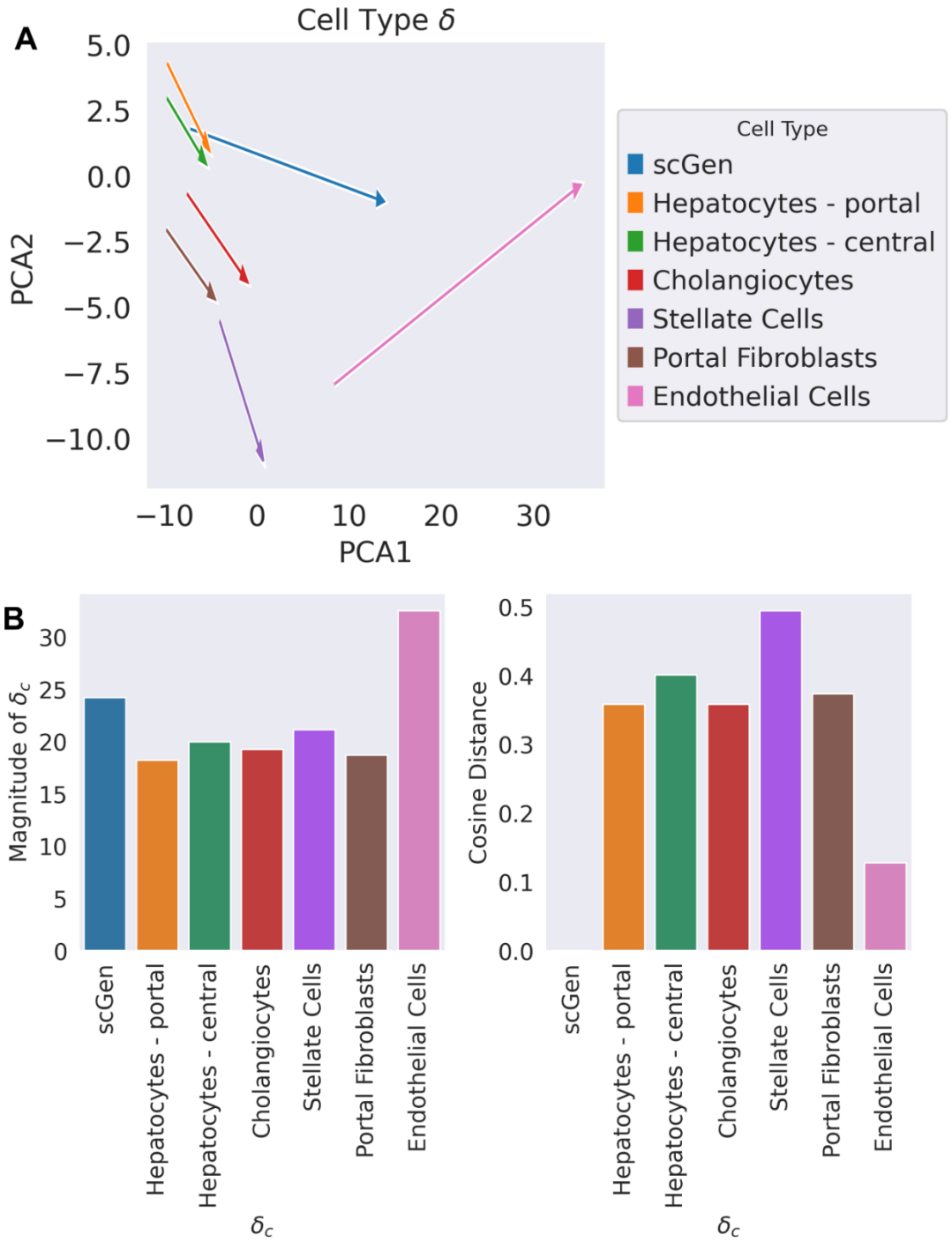

Figure 3.3 All δ_c *s* deviate from δ_{scgen} with respect to their direction across most cell types in the TCDD mouse liver snRNA-seq dataset. A) A PCA visualization of the calculated δc's for a VAE trained on all cell types. Each arrow represents the calculated δ for a particular cell type B) Bar plots of the magnitude of the δ and other individual δc's, and δc cosine distance from the δ.

Figure 3.3 (cont'd)

A cosine distance of 0 represents a δc in the same direction as δ, of 1 represents a δc orthogonal to δ, and of 2 represent a δc in the opposite direction as δ.

It is worth noting that outlier cell types could influence the calculation of δ_{scGen} since the latent space is highly sensitive to changes in the composition of the training set. Thus, if I were to take out endothelial cells and stellate cells from the analysis and retrain the model, it does not necessarily fix the issue of outliers influencing the relative magnitude and direction of δ_{scGen}. Instead, their removal exaggerates other differences in the latent space during the process of dimensionality reduction whereby cholangiocytes become a major outlier (Figure 3.4).

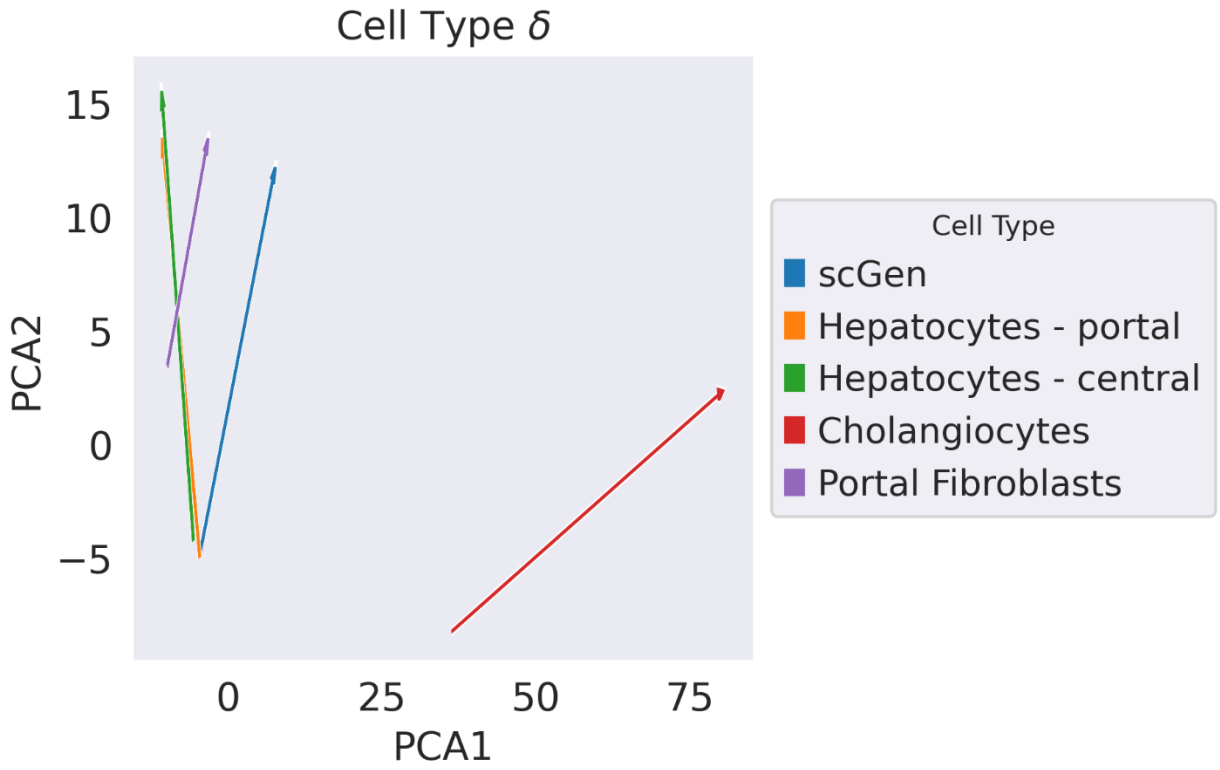

Figure 3.4 δ_c s deviation from δ with respect to their direction across most cell types is sensitive to training dataset. A PCA visualization of the calculated δ_c's for a VAE trained on all cell types excluding endothelial cells and stellate cells. Each arrow represents the δ_c associated with a specific cell type.

I conclude that a fixed model of δ cannot account for the full scope of heterogeneity that exists within many relevant toxicological datasets, and that the averaged δ_{scGen} is an inadequate way to

predict perturbations in the varied mouse liver cells. Instead of using a fixed δ_{scGen} I need a new δ prediction method that conditions on the transcriptomic state of the control cells. To do this, I introduce a regression-based correction to the data in order to more accurately capture δ_c.

3.2.3 scVIDR better predicts cell specific differences on the latent space of mouse liver TCDD perturbation than scGen.

When analyzing the latent space, I observed that the centroid of the control population for a particular cell type was predictive of δ_c ($R^2 \approx 0.9$), which led me to hypothesize that information about the perturbations of each cell type on the latent space might be encoded in the location of the corresponding centroid. I took a simple approach of using δ_c using the centroids of each control population as input data. From this point forward I will refer to this method as scVIDR.

scVIDR improved on scGen in terms of predicting cell type specific changes in gene expression (Figure 3.5 A). scVIDR's prediction, δ_{scVIDR}, of $\delta_{hepatocytes-portal}$ is closer to the PCA projection than scGen's prediction, δ_{scGen}. Additionally, scVIDR better predicts the magnitude and direction of $\delta_{hepatocytes-portal}$ than scGen. I conclude from these observations on the latent space that scVIDR is a more viable way to infer δ_c.

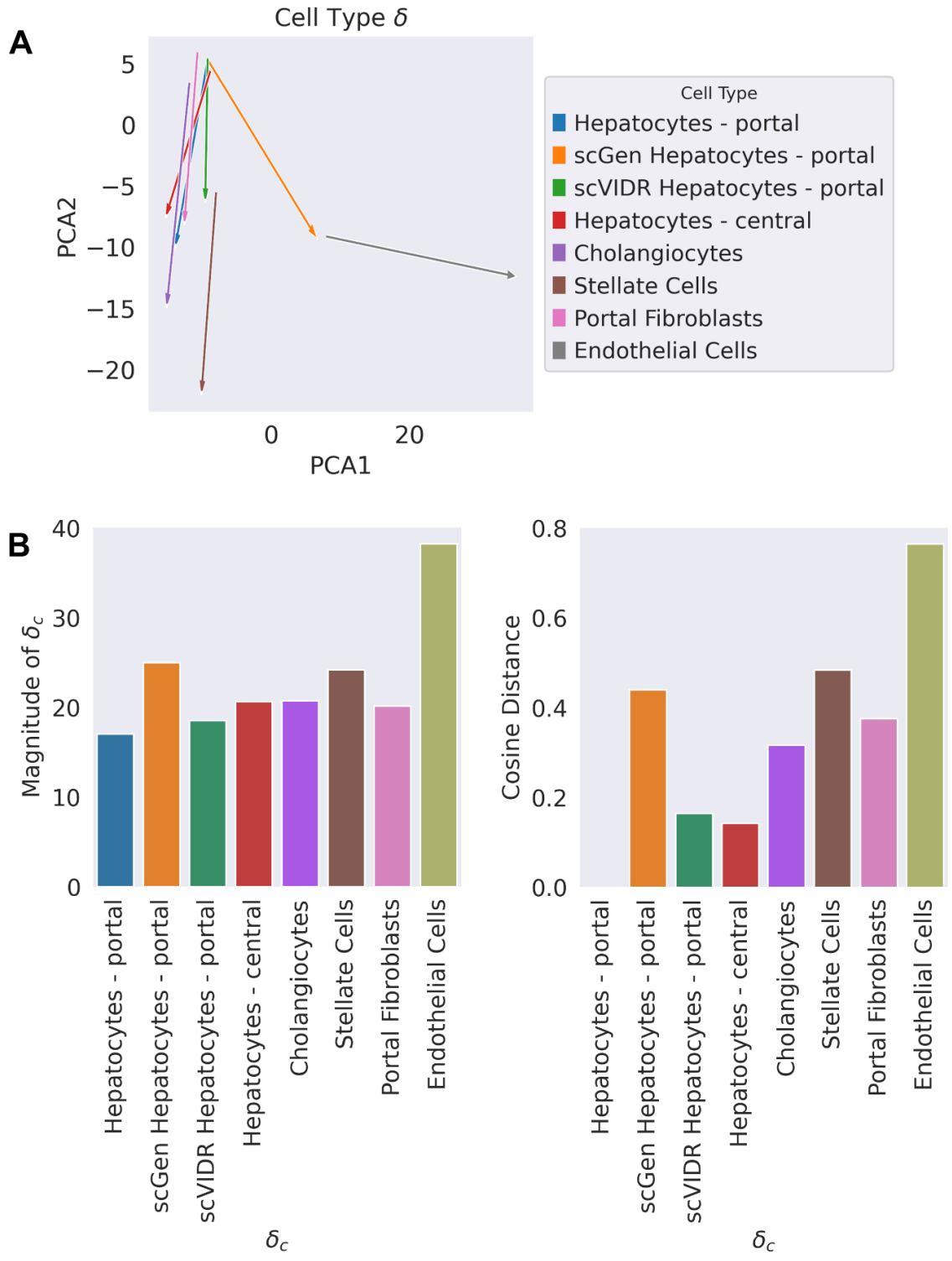

Figure 3.5 δ_{scGen} deviates more from $\delta_{Hepatocytes-portal}$ than δ_{scVIDR} . A) A PCA visualizationof the calculated δ_c's for a VAE trained without portal hepatocytes. "scGen Hepatocytes – portal" refers to the prediction by scGen (δ_{scGen}), and "scVIDR Hepatocytes –

Figure 3.5 (cont'd)

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 $\delta_{Hepatocytes-portal}$.

scVIDR is equivalent to scGen when there is only one cell type in the training dataset (Figure 3.6). This is because regression on the latent space of one cell type returns the δ of the entire dataset.

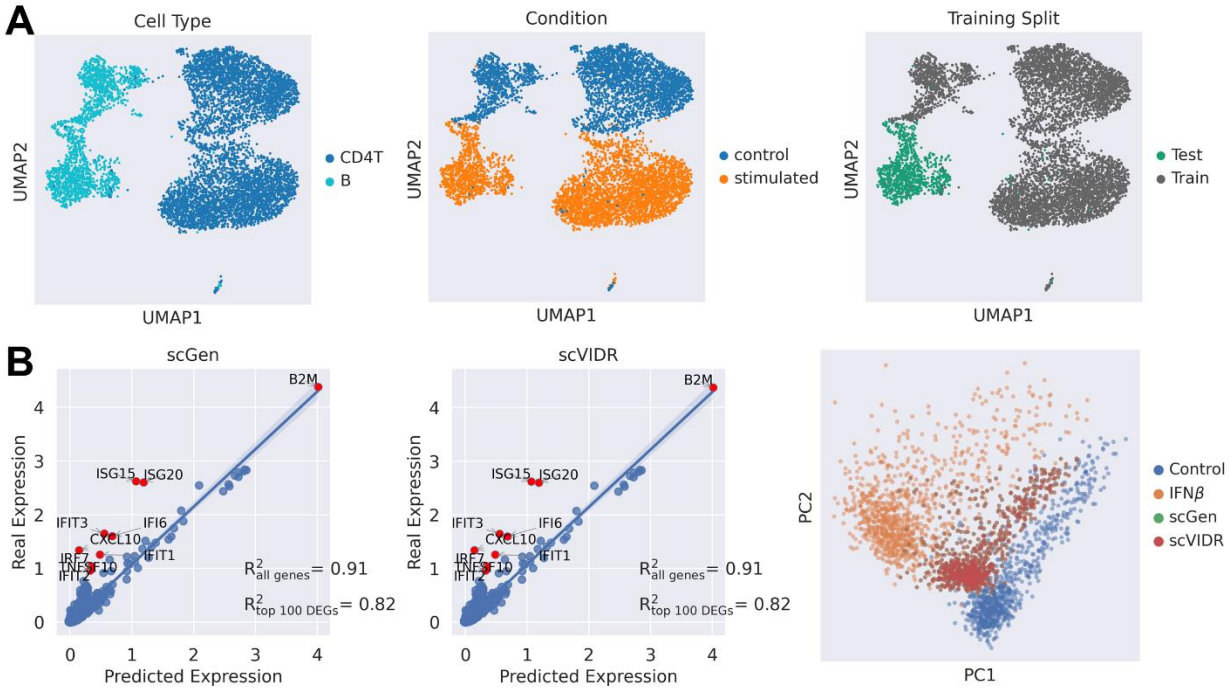

Figure 3.6 scVIDR is equivalent to scGen when training on a single cell type. A) A UMAP projection of latent space of single-cell expression of two cell types from Kang et al[99]: CD4T and B cells. They are colored by cell type, condition and train test split. B) Validation of prediction of B-cell perturbation when the 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 PCA plot of the predictions is displayed where each point represents an individual cell. Ground truth is in orange and the

Figure 3.6 (cont'd)

control data is colored blue. Predictions by scGen and scVIDR are colored green and red respectively.

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

I next evaluated scVIDR by comparing its ability to predict changes in gene expression in the mouse liver dataset with other state-of-the-art algorithms. Our training set (Figure 3.7A) consisted of all control and TCDD-treated cell types except for TCDD-treated portal hepatocytes which were used for model evaluation. I compared the performance of scGen[30], scPreGAN[98], CellOT[97], and scVIDR on the top 5000 highly variable genes (HVGs), and the top 100 differentially expressed genes (DEGs). When predicting the gene expression of portal hepatocytes, each method generated a set of virtual portal hepatocytes (Figure 3.7B). I 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 3.7C). Across DEGs, scVIDR produced an average R^2 of 0.81 (Figure 3.7C). Continuing the evaluation across all cell types (Figure 3.7D), leaving out one cell type perturbation at a time as described above for portal hepatocytes, our model outperformed all other models (with p-value < 0.001, one sided Mann-Whitney U Test) both when evaluated on HVGs and DEGs.

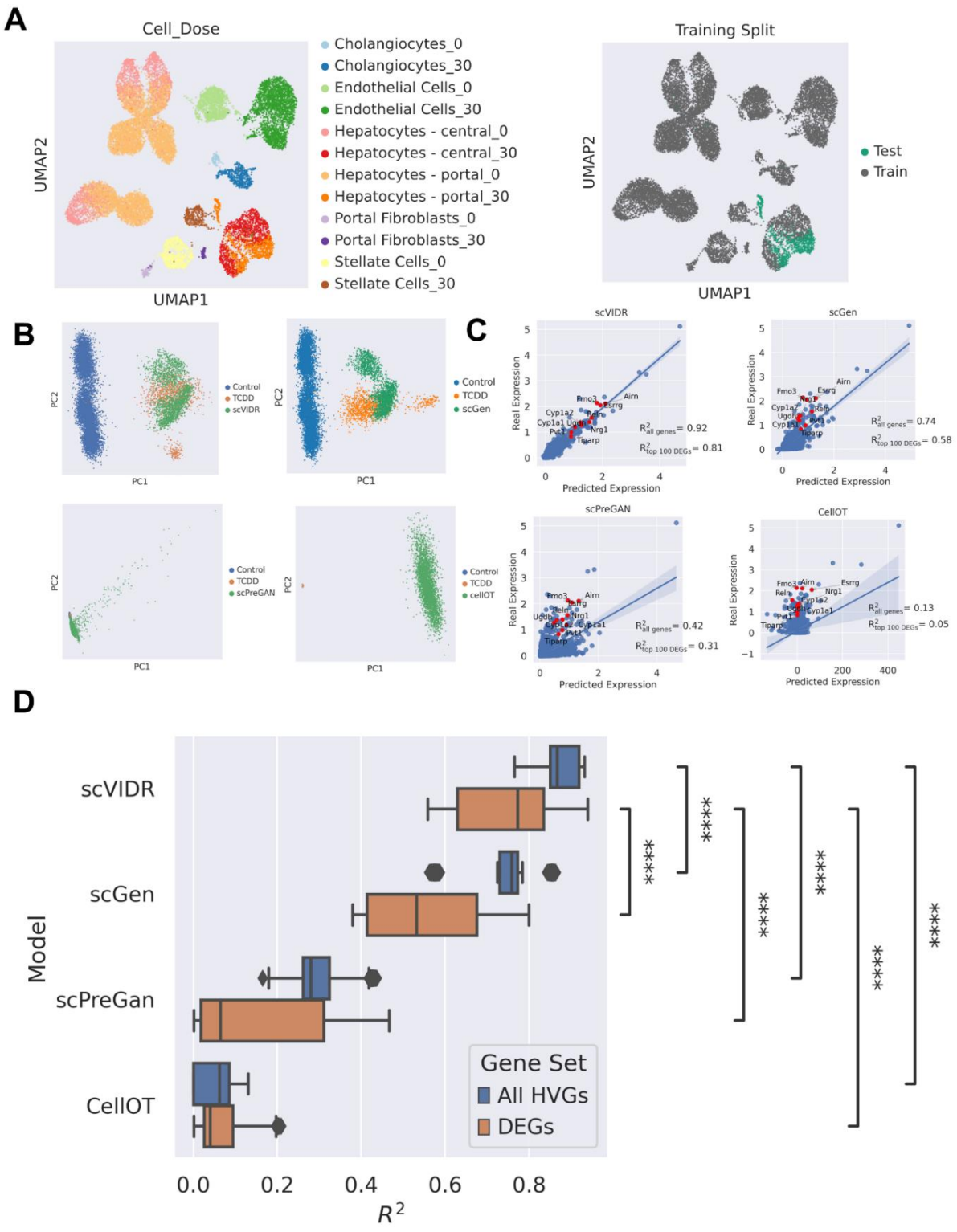

Figure 3.7 Prediction of in vivo single cell gene expression of portal hepatocytes from mice treated with 30 µg/kg of TCDD. A) UMAP projection of the latent space representation of scVIDR for control and treated single-cell gene expression. In A) each cell type and dose in

Figure 3.7 (cont'd)

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 hepatocytes responses following treatment with 30 μg/kg of TCDD using scGen[30], scVIDR, scPreGAN[98], and CellOT[97]. 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 for prediction across all liver cell types treated with 30 μg/kg of 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 one sided Mann-Whitney U test. **** indicates p-values < 0.001.

I further evaluated scVIDR on an additional dataset of IFNβ-treated PBMC [87]. A similar benchmark was performed as before (Figure 3.8), however instead of using the top 5000 highly variable genes, I used the top 6998 genes to be consistent across models compared. For this dataset, $t = 1$ labels PBMCs treated with 100 U/ml IFN-β , $t = 0$ for untreated PBMCs. The left-out cell type being predicted is B cells (Figure 3.7A). Across HVGs, the models yielded R^2 values of 0.97, 0.92, 0.77 and 0.66 and across DEGs, and R^2 values of 0.96, 0.86, 0.80, and 0.84 for scVIDR, scGen, scPreGAN, and CellOT respectively (Figure 3.7C). When accuracy was assessed for all cell types, scVIDR significantly outperformed all other models (Figure 3.8D).

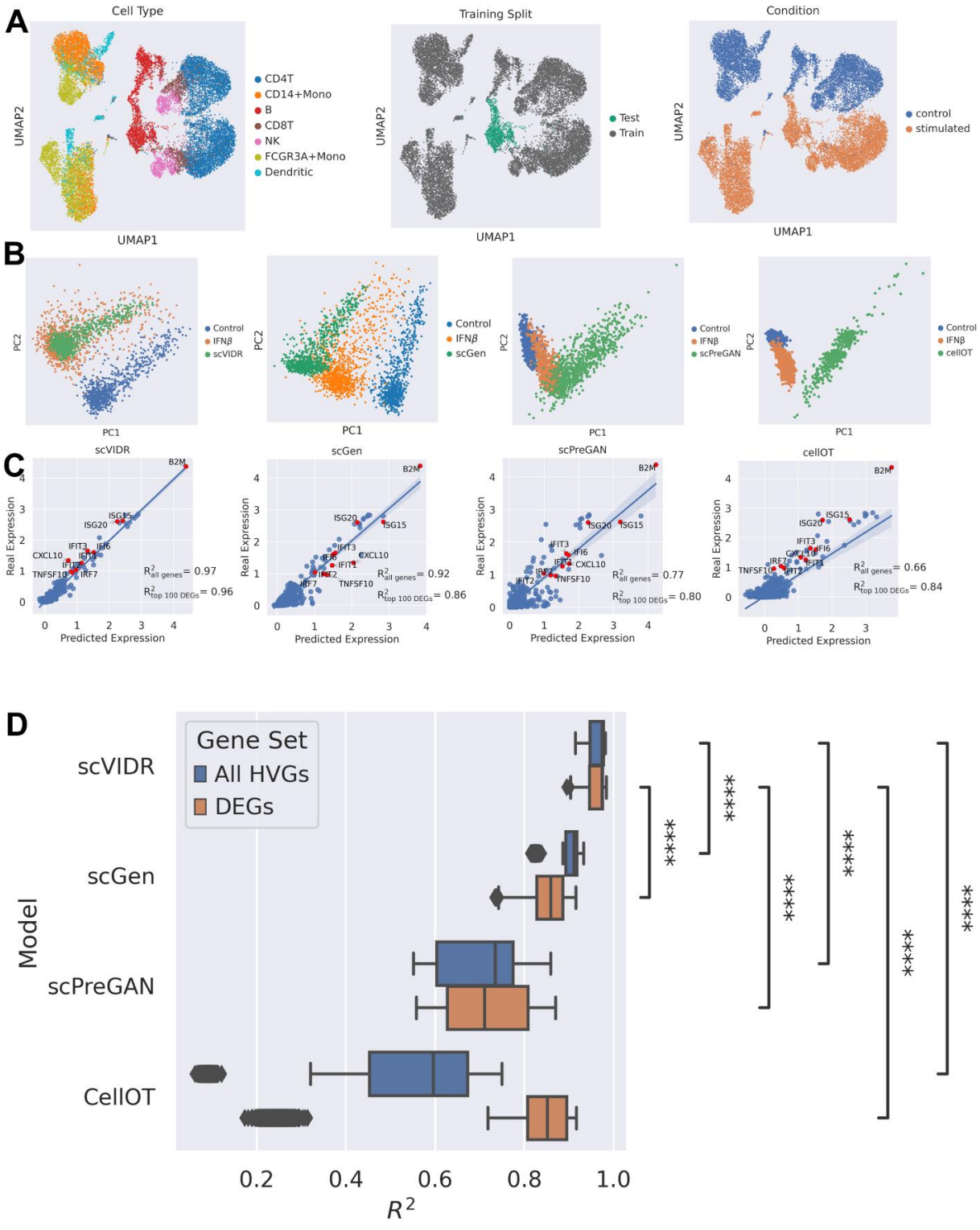

Figure 3.8 Prediction of in vitro response of B cells to IFNβ. A) UMAP projection of latent space of scVIDR for treated and untreated single-cell expression. UMAP plots are colored by

Figure 3.8 (cont'd)

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). Prediction performance distributions were compared using one sided Mann-Whitney U test. **** indicates p-values < 0.001.

To make sure that scVIDR is predicting unknown physiologies in chemical perturbations, I performed an additional experiment on IFNβ treated B-cells by selecting a subset of genes unique to IFNβ response of B-cells. To do this I take the set of DEGs unique to B-cells (Figure 3.9A). I show in Figure 3.9B that scVIDR consistently outperforms other models at predicting the gene expression means across all unique DEGs.

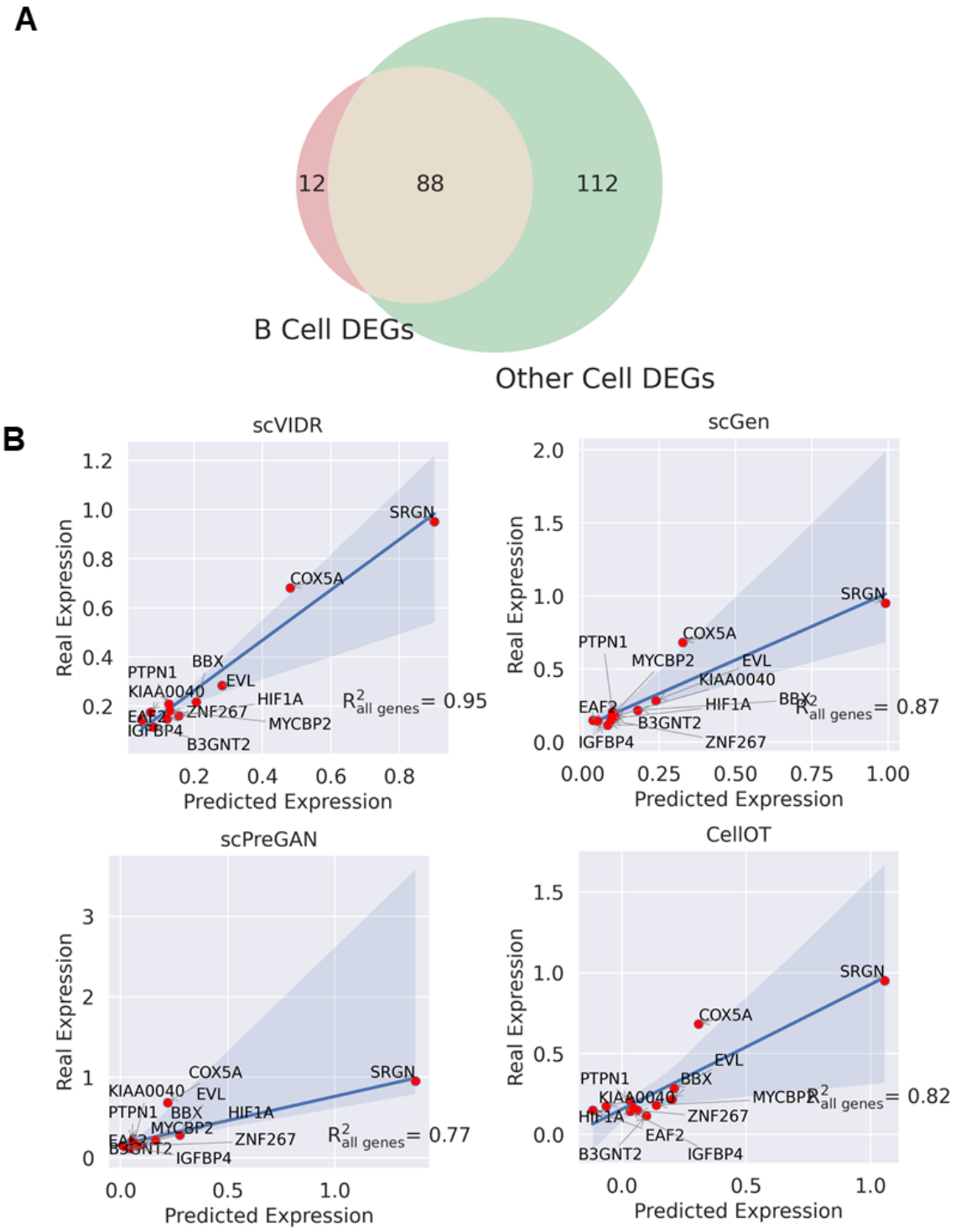

Figure 3.9 Evaluating scVIDR on differentially expressed genes unique to B-cells. A) A Venn diagram of the set of all top 100 differentially expressed genes (DEGs) for each cell type in the IFNβ treated PBMC dataset. Overlap between the top 100 differentially expressed gene in B-

Figure 3.9 (cont'd)

cells and the set of all top 100 differentially expressed genes in all other cell types is shown. B) Regression plots of predictions of the differentially expressed genes unique to B-cells (red portion of panel A). Each point represents the mean expression for a particular gene. Shaded region represents the 95% confidence interval.

3.2.5 scVIDR predicts rat phagocyte perturbation by LPS6 better than all other state of the art algorithms.

Next, I extend the idea of predicting chemical perturbations across cell types to cross-species predictions. The index c in δ_c now refers to the species that the cells came from rather than individual cell type. I also attempt to account for species differences in the same way that I try to correct for cell type differences using scVIDR.

I benchmark this potential for scVIDR on a dataset from Hagai et al[100], where mononuclear phagocytes were harvested from four different mammal species (rat, pig, mouse, and rabbit). $t = 1$ refers to phagocytes treated with 100 ng/ml IFN-β , and $t = 0$ for phagocytes treated with control, and the left-out species I want to predict are rats (Figure 3.10A). I see that scVIDR still outperforms all other algorithms at predicting perturbation of rat phagocytes (Figure 3.10B, C), with a correlation of $R^2 = 0.92$ for HVGs, and an $R^2 = 0.76$ for highly variable genes. I observe that while the predictions are more accurate for scVIDR, it still has difficulty predicting the

DEGs for rats.

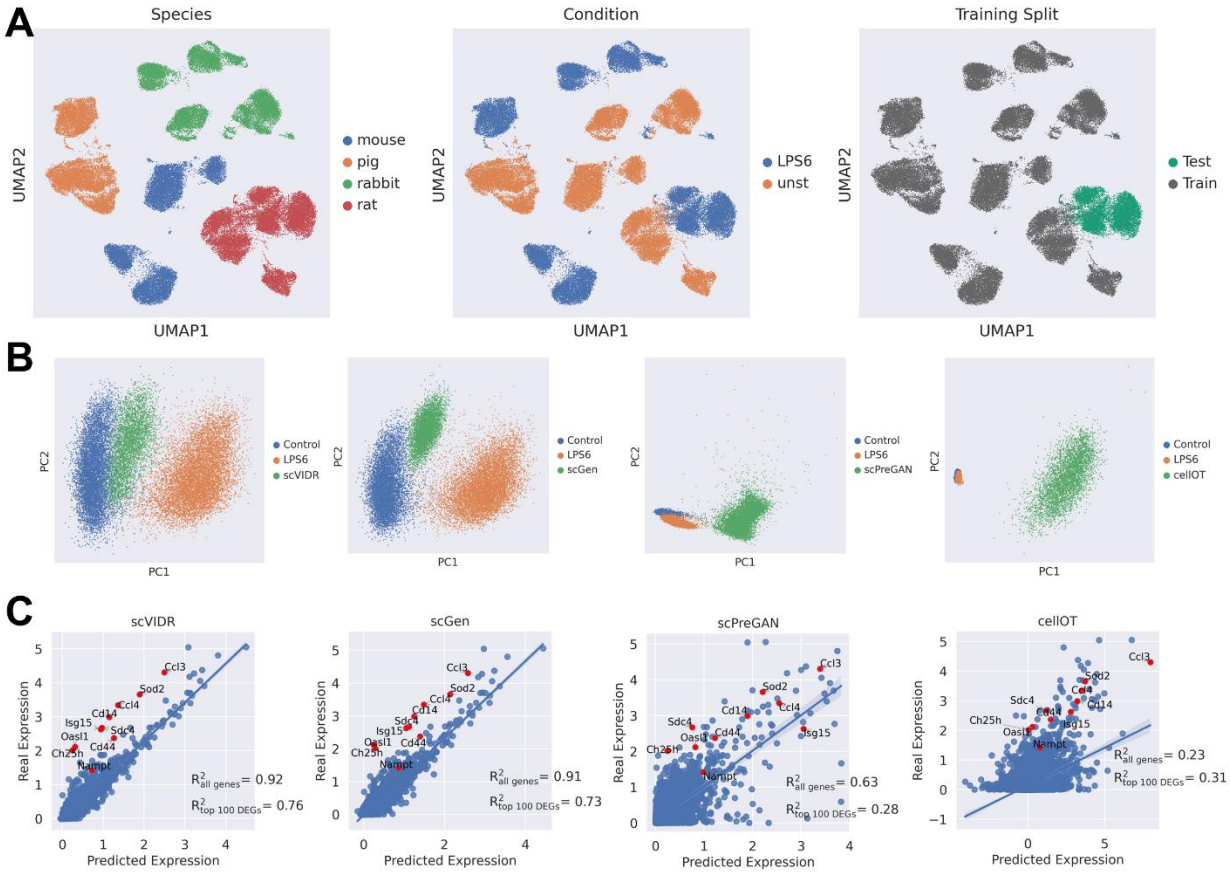

Figure 3.10 scVIDR predicts the effects of LPS6 on rat cells from mouse, rabbit, and pig cells better than other state-of-the-art algorithms. A) UMAP projection of latent space of scVIDR for treated and untreated single-cell gene expression. UMAP plots are colored by species, training split, and treatment condition, respectively. B) PCA plot of scGen, scVIDR, scPreGAN, and CellOT predictions of rat phagocytes 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.

3.3 Discussion

I have demonstrated several cases where scVIDR can predict single-dose chemical perturbations in scRNA-seq data. First, I establish that scGen is an inadequate algorithm for prediction of highly heterogenous tissues. Then I show how to improve scGen using regression to predict the

effect in a more cell-type specific way using an algorithm I developed, scVIDR. I then utilized scVIDR to predict chemical perturbations in three different scenarios. In the first scenario, I predict TCDD-induced perturbation in mouse liver cells. In the second where I predict IFNβ stimulation of PBMCs. In the third scenario, I predict LPS6-induced chemical perturbation of phagocytes derived from rats. In every case, scVIDR outperforms all other algorithms.

A question that arises is why scGen and scVIDR both perform better on the PBMC dataset (Kang et al)[99] than the rat phagocyte (Hagai et al)[100] and mouse liver datasets (Nault et al)[79]. I hypothesize that this is in part due to inter-cell-type overlap in the responses. I quantify this by looking at the overlap in the top 100 DEGs across all cells or species in the dataset. I observe there is much higher average overlap DEGs between the in the Kang et al dataset than in all other datasets (Figure 3.11). This makes intuitive sense given that datasets with similar cell-type responses will have more overlap and thus will be easier to predict for the VAE. I show that while this is an overall limitation in the VAE paradigm, scVIDR is more robust this limitation than scGen (see Sections 3.2.4 and 3.2.5).

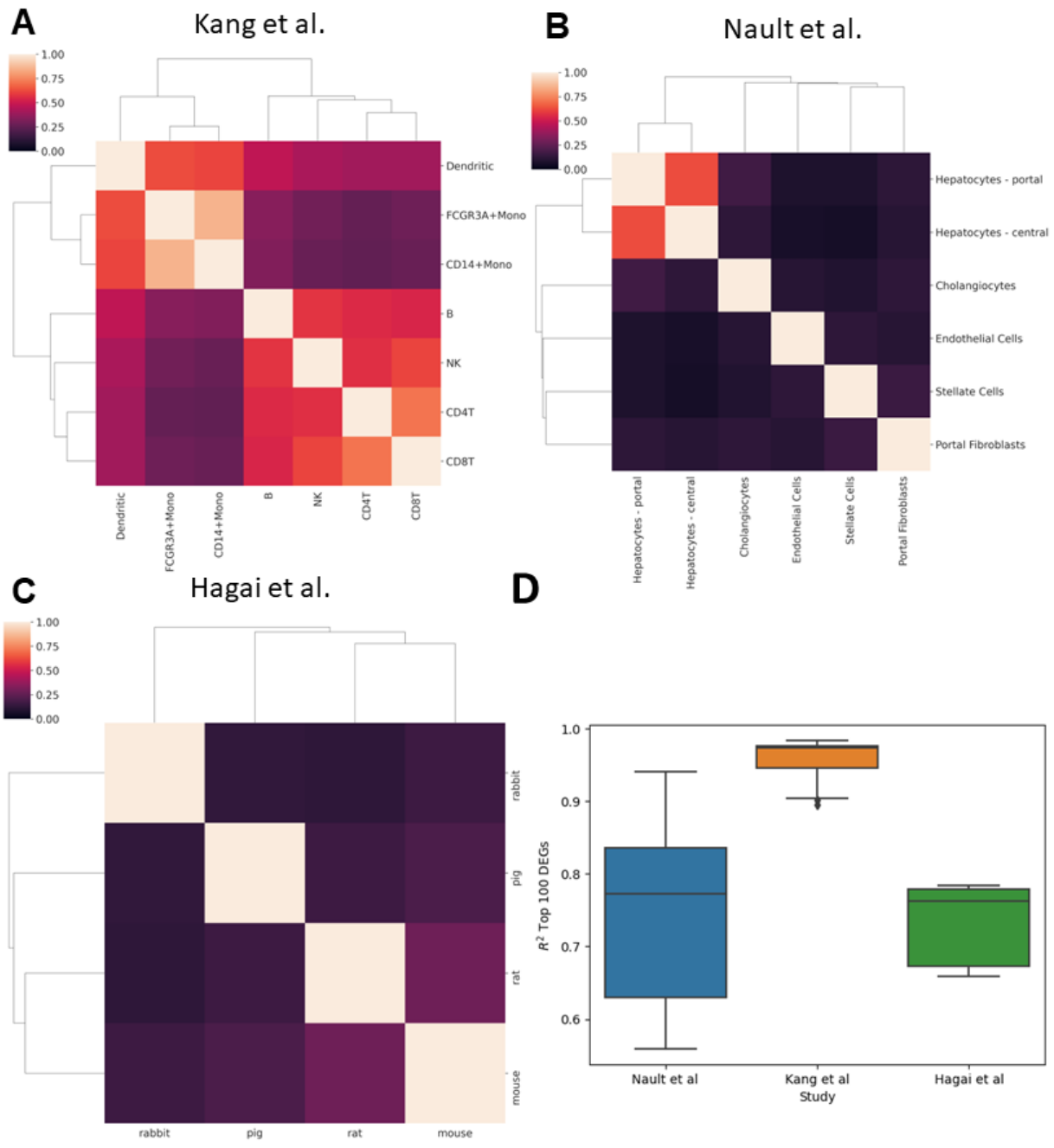

Figure 3.11 Overlap between differentially expressed genes reflects scVIDR prediction performance across datasets. A-C) Heatmaps of Jaccard similarity between top 100 DEGs for each pair of cells in each dataset. A) is the IFN-β treated PBMCs by Kang et al.[99] , B) TCDD Mouse liver from Nault et al.[79] C) is the LPS6 treated phagocytes by Hagai et al.[100] D) Boxplot of R^2 regression values for top 100 DEGs by scVIDR for each study in chapter 3.

However, DEG similarities do not completely explain why within datasets, there are differences in predictions between certain cell types. For example, the endothelial cell group, while having the smallest average overlap, has the second highest prediction scores in the Nault et al dataset (Figure 3.12A, D). Likewise, in the TCDD dataset, scVIDR performed better on the cell types most sensitive to TCDD, e.g., hepatocytes and endothelial cells (Figure 3.12B). When looking at cell types less sensitive to TCDD (e.g., cholangiocytes and stellate cells), the model often underestimated the expression of differentially expressed genes (Figure 3.12E). This is likely a result of a combination of factors including the similarity of the treatment to the control data (Figure 3.12B), smaller control cell populations (Figure 3.12C), and the overall low expression of highly variable genes (Figure 3.12E). Thus, I believe the VAE has less information to predict differential gene expression for these cell types. Additional control data for more rare subpopulations could help alleviate the prediction inaccuracy.

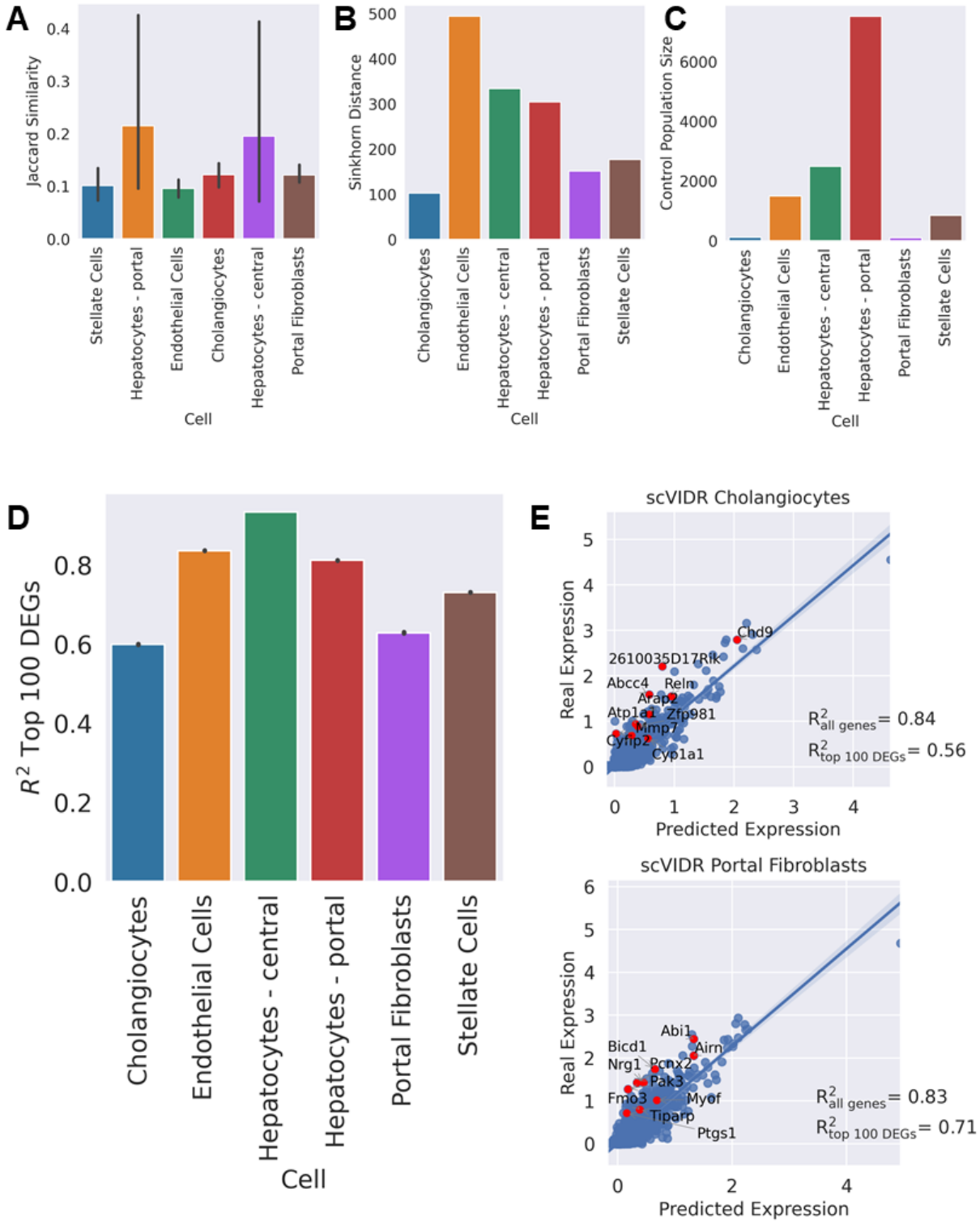

Figure 3.12 Impact of differential expression overlap, latent perturbation magnitude, control population size on overall model performance. A) Average Jaccard similarity for each

Figure 3.12 (cont'd)

cell type in the Nault et al study[79] B) Sinkhorn distance between the latent distributions of the control and 30 μg/kg doses of TCDD of each cell type on the latent space. C) Bar plot of the control group cell population size for each cell type. D) Bar plot of mean gene R^2 for each individual cell type when predicting only the 30 μg/kg dose of TCDD. 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. Shaded region around regression line represents the 95% confidence interval.

Mapping the combinatorial space of chemical perturbation is becoming more and more crucial as industrial practices and drug development accelerate the number of chemicals toxicologists need to evaluate for safety. Doing so experimentally is prohibitively expensive due to the sheer size of the chemical x dose perturbation space. Computational algorithms that can support current endeavors to do high throughput testing of chemical perturbations are needed now more than ever. I envision a future where technologies such as scVIDR can be applied to large single-cell chemical perturbation atlases to reduce the number of experiments needed to evaluate a drug or chemical. Not only does this represent a practical endpoint to technologies like scVIDR, but also an ethical one, as reduction of animal testing has become a high priority in the field of toxicology. Thus, I feel that this work is of immense significance to the fields of pharmacology and toxicology.

3.4 Methods

3.4.1 A mathematical description of variational autoencoders

Variational Autoencoders[28] aim to estimate the posterior probability function, $p_\theta(z|X)$ of a latent process, z, given a set of observations, X. By Bayes theorem I can calculate the posterior probability:

$$p_\theta(z|X) = \frac{p(X,z)}{p(X)}$$

However, calculating the posterior in this way is intractable due to the difficulty in computing the marginal distribution:

$$p(X) = \int_z p(X|z)p(z)dz$$

Thus, I instead aim to approximate $p_\theta(z|X)$ by minimizing the Kullback-Leibler (KL) divergence between $p_\theta(z|X)$ and some gaussian distribution, $q_\phi(z|X)$, whose parameters, ϕ, are calculated by a neural network. This neural network is termed the encoder.

I can calculate the KL divergence as:

$$D_{KL}(q_\phi(z|X)||p_\theta(z|X)) \ = E_{z\sim q}\left(log\frac{q_\phi(z|X)}{p_\theta(z|X)}\right)$$

Substituting $p_\theta(z|X)$ with Bayes theorem and rearranging the terms:

$$D_{KL}(q_\phi(z|X)||p_\theta(z|X)) \ = \log p_\theta(X) - \ E_{z\sim q}\left(\log\frac{p_\theta(X,z)}{q_\phi(z|X)}\right)$$

The second term in the equation above is also known as the evidence lower bound, or ELBO. Since the KL divergence must be a positive value, I can minimize it by maximizing the ELBO. I can rewrite the ELBO as:

$$E_{z\sim q}\left(\log\frac{p(X,z)}{q_\phi(z|X)}\right) = \ E_{z\sim q}\left(\log\frac{p_\theta(X|z)p_\theta(x)}{q_\phi(z|X)}\right) = E_{z\sim q}(\log p_\theta(X|z)) - E_{z\sim q}\left(\log\frac{q_\phi(z|X)}{p_\theta(z)}\right)$$

Since the second term in the equation above is equivalent to the definition of the KL Divergence I can rewrite the ELBO as:

$$E_{z\sim q}(\log p_\theta(X|z)) - D_{KL}(q_\phi(z|X)||p_\theta(z))$$

The first term is known as the reconstruction error. This maximizes the likelihood that I will generate values from the latent space that match our observations. The second term is the KL divergence term, which is predicated by the KL divergence between the distribution estimated by the encoder and the prior distribution which is a standard normal multivariate distribution. This second term encourages structure in $p(z)$ as minimizing the difference between $q_\phi(z|X)$ and, as a result, maximizing disentanglement in $p(z)$. I can now construct our objective function, $L(\theta,\phi)$, as the following:

$$L(\theta,\phi) = \ -E_{z\sim q}(\log p_\theta(X|z)) + D_{KL}(q_\phi(z|X)||p_\theta(z))$$

Unfortunately, naively trying to take the gradient with respect to ϕ to minimize this function will result in highly variable gradients and thus variable training results. To fix this I instead utilize the reparameterization trick, where I sample from the latent space such that z is deterministic with respect to some noise variable, $\epsilon \sim N(0,I)$, or:

$$z_i = \mu_\phi + \Sigma_\phi^{\frac{1}{2}} \cdot \epsilon$$

Where, μ is the mean vector and Σ is the covariance matrix of the inferred distribution.

3.4.2 Single cell expression datasets and preprocessing for scVIDR.

Data from Nault et al[79] was collected and processed 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, I select the top 5000 most highly variable genes to do our analysis on. The preprocessing was carried out using the *scanpy.pp* package using the *normalize_total*, *log1p*, and *highly_variable* functions[87].

The Nault of et al dataset[79] comprised of single nuclei RNA-seq of C57BL6 of flash frozen mouse livers. Mice in this dataset were administered, sub-chronically, 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[79]. Thus, there is a small size for the immune cell populations in the low-dose datasets versus the high-dose. PBMC data from Kang et al[99], Study B data from Zheng et al[16], and species data from Hagai et al[100],was accessed as a processed dataset from Lotfollahi et al[30].

When training scGen and scVIDR, batch effects are accounted for with the *scvi.data* package using the *setup_anndata* function[29]. Differential abundances of cells in different groups are accounted for by random sampling with replacement the same number of cells for each dose and random sampling without replacement the same number of cells for each cell type.

3.4.3 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[30] which in turn is buil on the python pachakge scVI v. 0.13.0[29]. I extend existing code bases to include linear regression on the latent space.

Hyperparameters for the model and training are the default values selected by scGen v. 2.0.0. Table 3.1 outlines the model hyperparameters used in deploying scVIDR and scGen:

Hyperparameter	*Value*
Latent dimension	100
Number of layers	2
Layer width	800
Dropout rate	0.2
Kullback-Leibler weight	5 $* 10^{-5}$

Table 3.1 Hyperparameters for scVIDR's and scGen's variational autoencoder model.

Table 3.2 outlines the training hyperparameters when deploying scVIDR and scGen:

Hyperparameter	*Value*
Training epochs	100
Learning rate	0.001
Learning rate decay	10^{-6}
Optimizer	Adam
Optimizer epsilon	0.01
Early stopping	True
Early stopping patience	25

Table 3.2 Hyperparameters for scVIDR's and scGen's training scheme.

3.4.4 scVIDR calculation of $\hat{\delta}_c$

If I 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, I can approximate a function based on $\bar{z}_{c=B,p=0}$, or:

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

Where I 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[101].

CHAPTER 4
PREDICTION OF MULTIPLE DOSE CHEMICAL PERTURBATIONS ACROSS CELL STATES USING VARIATIONAL AUTO-ENCODERS

4.1 Introduction

In this chapter, I will examine the ability of models such as scVIDR to not only predict single dose chemical perturbations, but to interpolate response across unknown doses in the latent space. Here, as in the previous chapter, I have some unknown cell type C whose response to a chemical is unknown. However, unlike the previous chapter, I am not only looking at a fixed dose of a particular chemical, but rather the spectrum of response that results from treatments at different levels of concentration. I refer to the cellular response to a specific concentration of a given chemical as the dose-response curve for that particular cell type. I will refer to the prediction of the response to multiple doses of a chemical as dose-response prediction.

More formally, given that I know the dose-response for some cell types A and B, can I predict the dose-response of C? I perform this task by extending the scVIDR model from chapter 3 to account for multiple doses. To do this I take the latent δ_c and interpolate on it log-linearly to predict doses between the lowest and highest measured in the dataset (Figure 4.1).

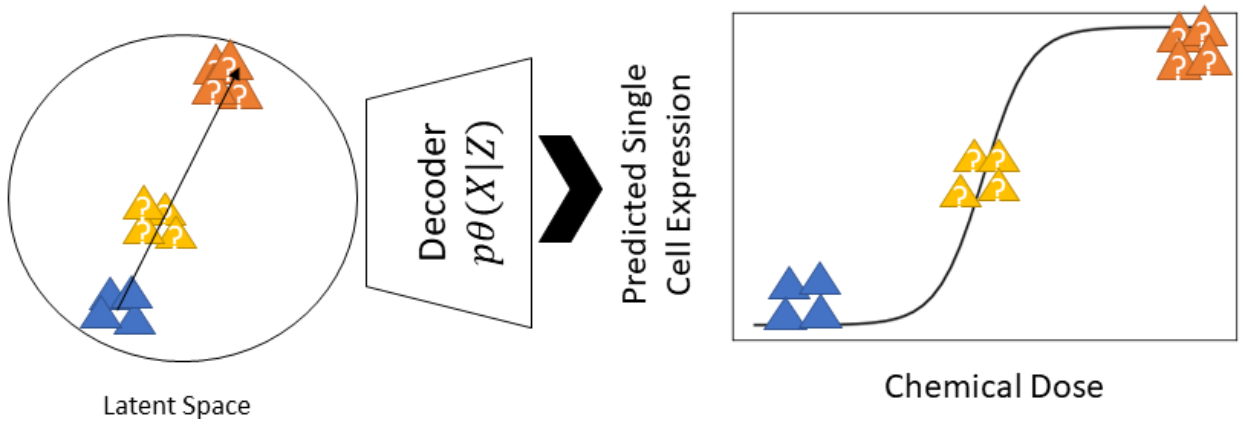

Figure 4.1 Schematic of perturbation prediction across multiple doses in scVIDR. Outline of *scVIDR* for prediction of the unknown response to multiple doses for some unknown cell type. Log-linear interpolation on δ is used to predict dose dependent changes in gene expression in the latent space. The latent space representations are then projected back into gene expression space using the decoder.

To my knowledge there is only one other model in the literature that performs this particular task, the compositional perturbation autoencoder (CPA)[41]. CPA creates multiple autoencoders that the user can add together in a modular way to remove variation for multiple covariates (e.g., dose, time-point, cell type, etc.). As a result, the user can train a model on multiple covariates at the same time to create latent representations of each one. However, CPA requires a large amount of training data in order to create accurate latent spaces for each covariate. Our model, by comparison, requires much less data to make accurate predictions. In addition, CPA is based on a regular autoencoder framework, which unlike scVIDR's VAE, encodes sparse representations of the latent space (i.e., the space between data points does not encode much information)[102]. This makes them less amenable to interpolation. VAEs by comparison learn using probability distributions, and better describe continuous trajectories in the data[28].

I evaluate the abilities of this model on two datasets. The first dataset is an extension of the TCDD mouse liver dataset described before, where instead of a single fixed dose of 30 μg/kg TCDD, there are seven additional doses ranging from 10 ng/kg to 10 μg/kg TCDD (Nault et al)[86]. I show that simply by interpolating on δ_c I can predict the unknown dose-responses for a wide range of chemicals. The second dataset is the Sciplex dataset (Srivatsan et al)[95], which is comprised of three cancer cell lines (A549, MCF7, and K562) treated with four doses of 188 drugs. I show that I can use scVIDR to predict the dose-response of thirty-seven different chemicals.

In addition to the prediction of the response to multiple doses, I would also like to understand what genes are important in predicting the response for a particular drug. This would help validate the model when performing predictions. To do this I take advantage of the structure of the VAE to identify genes with the highest contribution to the dose-response according to scVIDR.

Finally, when calculating the dose-response in single cell data, I encounter heterogeneity in individual cell responses to the same dose of a particular chemical, either due to the inherent stochasticity in transcription[103], the chemical environment of the cell, or the internal state of the cell[8,9]. As a result, I would like a measure of how much a particular cell has responded transcriptionally to a chemical. I call the metric I have developed "pseudo-dose". I show that pseudo-dose accurately describes the variation in the response to TCDD exhibited by hepatocytes across the liver lobule.

4.2 Results

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

In this section I predict the response of liver cells to multiple doses of TCDD. Here, t is equal to the magnitude of the perturbation, which in my case is equivalent to the dose. Thus, $t = 0$ represents expression at dose 0 and $t = 30$ represents expression at dose 30, where the dose is in units of μg/kg for the Nault et al dataset[86]. As with the single-dose case, I train the model on the dose-response data for all cell types except one, for which only the $t = 0$ condition is kept. I calculate the $\hat{\delta}_c$ (equivalent to δ_{scVIDR} in section 3.2.2; see section 3.4.4) 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[30], I log-linearly interpolate on δ_{scGen}. Finally, those latent space representations are decoded back into gene expression space using the decoder portion of each of the models.

I analyzed a mouse liver snRNA-seq from the Nault et al dataset[86] 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 4.2 A). scVIDR outperforms scGen in approximating expression across the dose-response of TCDD in mouse liver. I used the mean R^2 score across all evaluated genes as my performance metric (Figure 4.2B). scVIDR significantly out-performed 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 4.2C). 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 4.2D). When predicting on just the DEGs, scVIDR significantly outperformed scGen for doses > 0.3 μg/kg (Figure 4.2E).

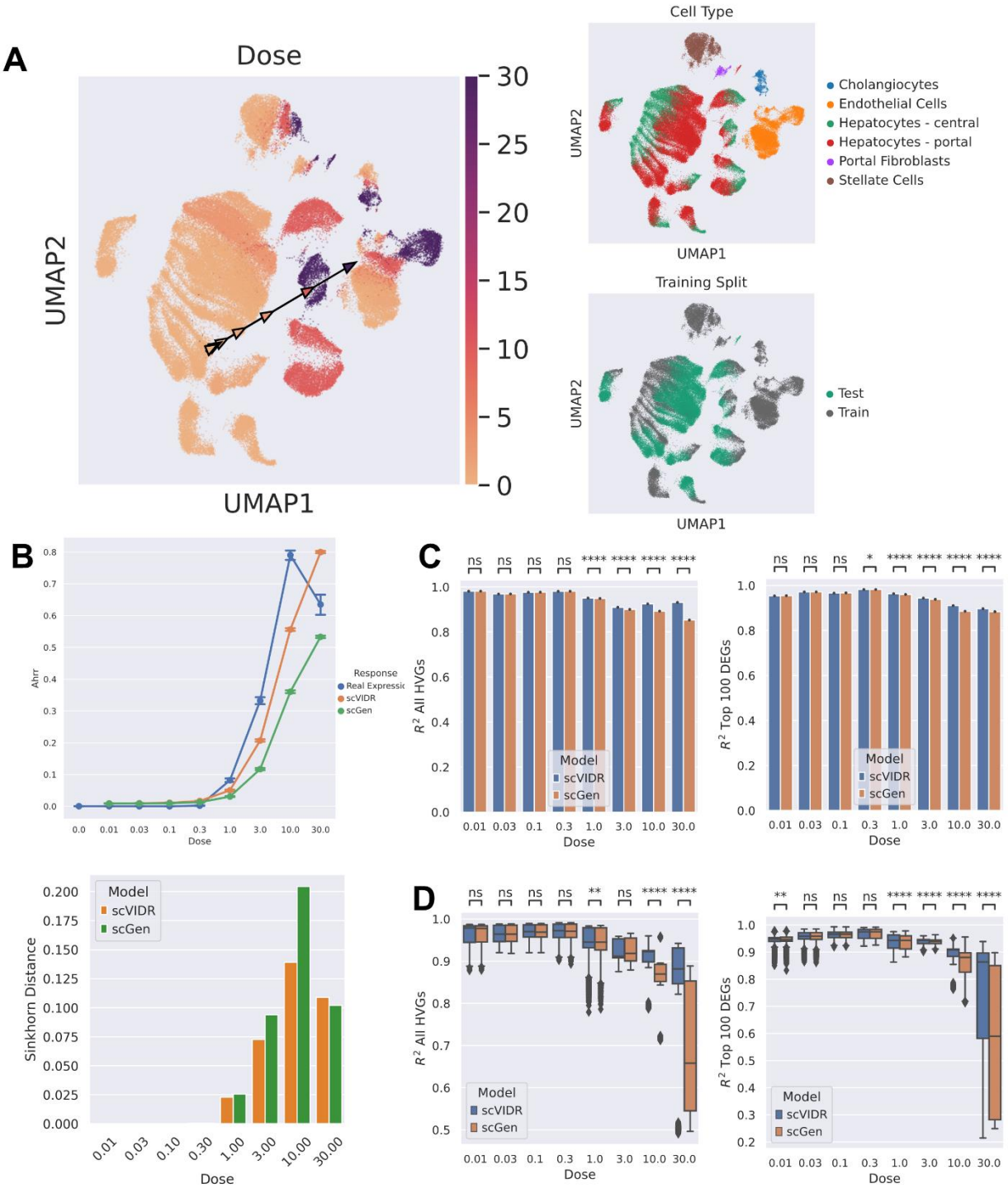

Figure 4.2 Prediction of in vivo singe cell expression of the TCDD dose-response in portal hepatocytes from mouse liver tissue. A) UMAP projection for the latent space representation of single cell expression across TCDD dose-response. Cells are colored by dose (μg/kg), cell type, and training split. Arrows on UMAP represent a δ calculated on UMAP space, with each

Figure 4.2 (cont'd)

arrowhead representing a specific dose denoted by its color. B) Dose-response prediction for Ahrr using scVIDR, and scGen. The differences between the predicted and true distribution of Ahrr at each dose are measured via the Sinkhorn distance. C) Bar plots of the R^2 of the gene expression means in portal hepatocytes for all highly variable genes and the top 100 differentially expressed genes. D) Box plot of the distribution of R^2 scores across all cell types in liver tissue.

I used scVIDR to predict the effects of a test set of 37 drugs out of 188 treatments in the sci-Plex dose-response data[95] at 24 hours for A549 cells (Figure 4.3). 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 4.3A). The dose-response for the 37 drugs was predicted as above by first calculating the $\hat{\delta}_{A549}$ between the control and highest dose for a particular drug and log linearly interpolating along the $\hat{\delta}_{A549}$ in order to predict the intermediate doses. I evaluated predictions made by scVIDR at the gene, drug, and drug pathway level. For the drug Belinostat, a histone deacetylase inhibitor, scVIDR improves on predictions of differentially expressed genes such as *MALAT1* relative to scGen (Figure 4.3B). When predicting gene expression of the DEGs in Belinostat treated A549 cells, scVIDR also significantly outperformed scGen on all doses (Figure 4.3C). On predicting the DEGs of all drugs with the same mode of action as Belinostat (Epigenetics), scVIDR similarly outperformed scGen on all doses (Figure 4.3D). Finally, when looking across all 37 drugs in the test dataset, I was able to predict the expression of DEGs significantly better than scGen on average for the 3 highest doses of 100, 1,000, 10,000 nM (Figure 4.3E).

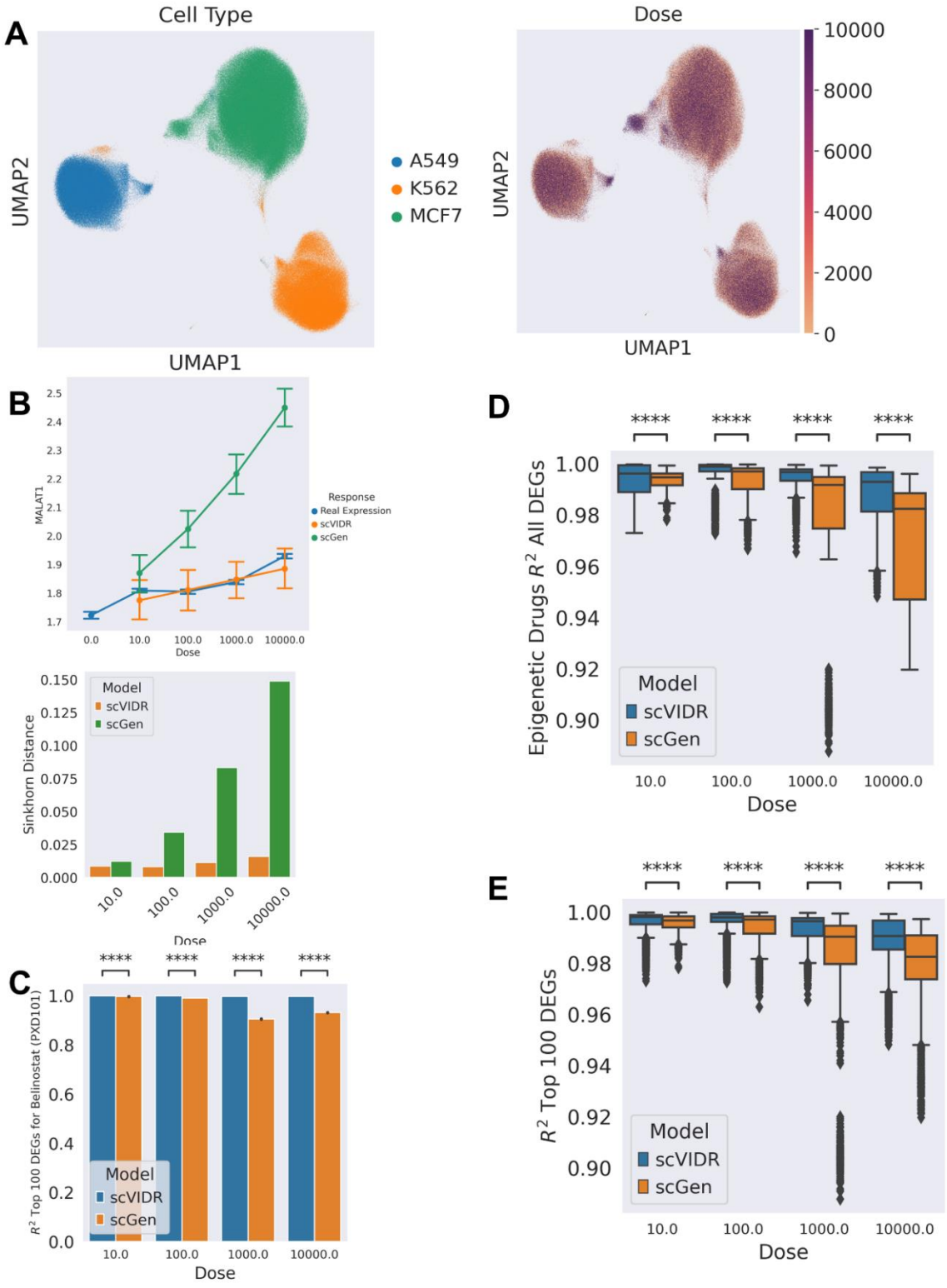

Figure 4.3 Prediction of in vitro dose-response of A549 cells to different drug treatments. A) UMAP of the latent space of single-cell expression colored by cell type and dose (nM)

Figure 4.3 (cont'd)

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.

4.2.2 More sophisticated models of dose-response interpolation do not improve cross cell type prediction when compared to log-linear interpolation.

The dose-response is described by a sigmoid relationship between the concentration of the chemical of interest and the measured physiological response. Often when describing this relationship, pharmacologists and toxicologists use the Hill equation to model this relationship. However, as the Hill model has parameters that are usually unknown to us, I must fit the model to a dataset. As a result of the lack of information on parameters (e.g., EC50) for specific drugs in specific cell types, how well the model matches the true biology is difficult to validate. The same though process can be extended to a threshold model of response, where the sigmoid relationship is described by a log-linear response that starts at a specified dose rather than 0. More concretely, I can evaluate each of the interpolation functions based on whether they can predict intermediate doses more or less accurately across several drugs. I performed the following experiment, where I took five random drugs from Srivatsan et al dataset (TGX-221, Crizotinib (PF-02341066), Tranylcypromine (2-PCPA) HCl, XAV-939, and Decitabine) and predicted their dose-response in A549 cells using three different interpolation methods: Hill, threshold, and log-linear (explanation of models in section 4.4.3). I show in figure 4.4 that there is no statistical difference between the log-linear and either the threshold or the Hill model of interpolation. This suggests that calculation of δ has a much higher impact on the prediction accuracy than how I interpolate on said δ.

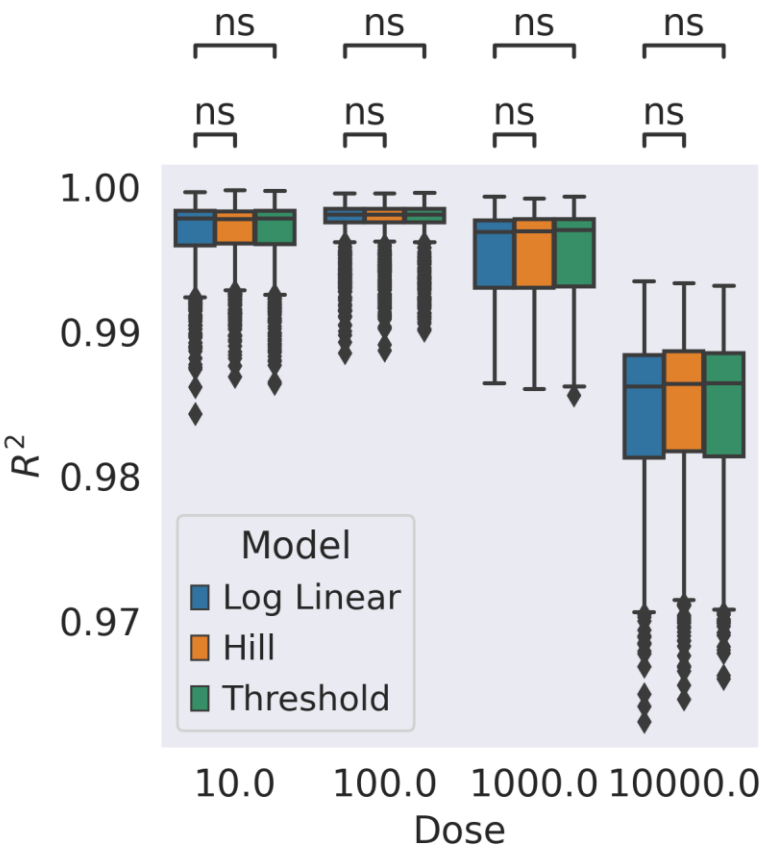

Figure 4.4 Comparison of dose-response predictions for five drugs in A549 cells. Boxplot of the distribution of predictions made for five drugs (TGX-221, Crizotinib (PF-02341066), Tranylcypromine (2-PCPA) HCl, XAV-939, and Decitabine) in A549 cells at four different doses (nM). Log-linear model predictions of the dose-response are in blue, Hill models of the dose-response are in orange, and threshold models of the dose-response are in green.

4.2.3 Regression on the latent space infers the relationship between predicted gene expression and $\widehat{\boldsymbol{\delta}}_{\boldsymbol{c}}$

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 scVIDR's predictions, I approximate the function of the decoder with linear regression (see section 4.4.4 for more extensive explanation). I take inspiration from the use of principal component analysis (PCA) in scSeq[104] and the development of the linear decoded variational autoencoders (LDVAE)[38]. 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 $m\ x\ g$ where m is the number of latent variables, and g is the number of genes. I 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 which, 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[38].

However, interpretability comes at the expense of model accuracy. LDVAEs have higher reconstruction error than standard VAEs on single cell data[38]. Similarly, using PCA and vector arithmetic to predict scSeq perturbations performed poorly compared to scGen[30]. As a result, one would like to try and interpret the latent space of a standard VAE. I present an approach to interpret the VAE's latent space using sparse regression.

I take an alternative approach to LDVAEs in which I instead approximate the non-linear function of the decoder in a standard VAE using sparse linear regression (Figure 4.5A). Sparse regression methods like LIME have been used to interpret complex models[105]. I specifically use sparse linear ridge regression, given that each gene has a non-zero contribution to each latent variable and gene weights are distributed parsimoniously. This gives us a linear transformation matrix, $\widehat{W}_{vae}$, that approximates the function of the decoder.

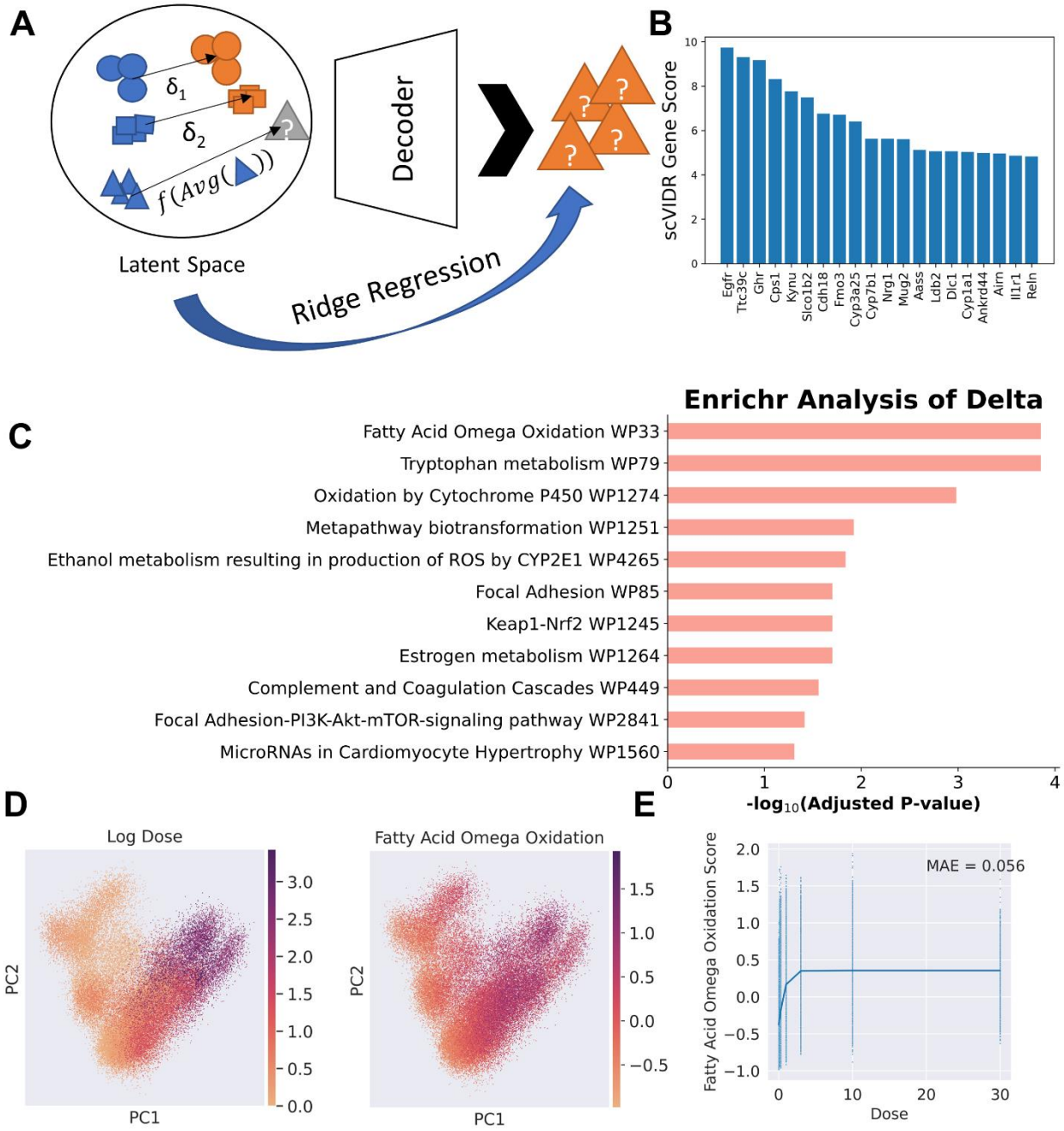

Figure 4.5 Interrogation of VAE using ridge regression in portal hepatocyte prediction. A) Schematic of calculating latent dimension weights using ridge regression. B) Bar plot of top 20 genes with the highest scVIDR genes scores. C) Enrichr analysis of the top 100 genes with respect to the scVIDR gene 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

Figure 4.5 (cont'd)

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

I 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:

$$Gene\ Scores = \hat{\delta}_c^T \widehat{W}_{vae}$$

In practice, I found that normalizing the weight matrix by its L2 norm gives better insights when interpreting the model (see section 4.4.4). Gene scores represent how significant changes in latent space dimensions will impact the decoded transcriptomic response when I 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 I increase the dose of my prediction by scVIDR.

I utilize a trained scVIDR model where portal hepatocytes were left out of training and the $\hat{\delta}_{c=Portal\ Hepatocytes}$ was approximated (Figure 4.5 B-D). Gene scores for $\hat{\delta}_{c=Portal\ Hepatocytes}$ 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 4.5B)[106]. To see whether this relationship extended to pathways involved in TCDD-induced genetic response, I performed Enrichr analysis[83,92] using the 2019 WikiPathways database[107] on genes with the top 100 gene scores (Figure 4.5C). Among the top enriched terms, I found the hallmarks of hepatic response to TCDD in mice, such as oxidation by cytochrome P450[108], fatty acid omega oxidation[109], and tryptophan metabolism[110]. 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 was used in calculating enrichment scores for each cell using Scanpy[87]. A sigmoid function was fit to the median enrichment score in each dose (section 4.4.6). I observed a small mean absolute error in my model and thus concluded that there was a sigmoidal dose-response relationship for the gene set generated by Enrichr (Figure 4.5 D, E).

4.2.4 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 Slingshot[111] and Monocle[112]. In pharmacology and toxicology, I experience a similar problem as cells of the same type have variable sensitivities to the same toxicant. Hence, I propose to order cells in terms of a latent dose. I call this ordering of cells a “pseudo-dose”.

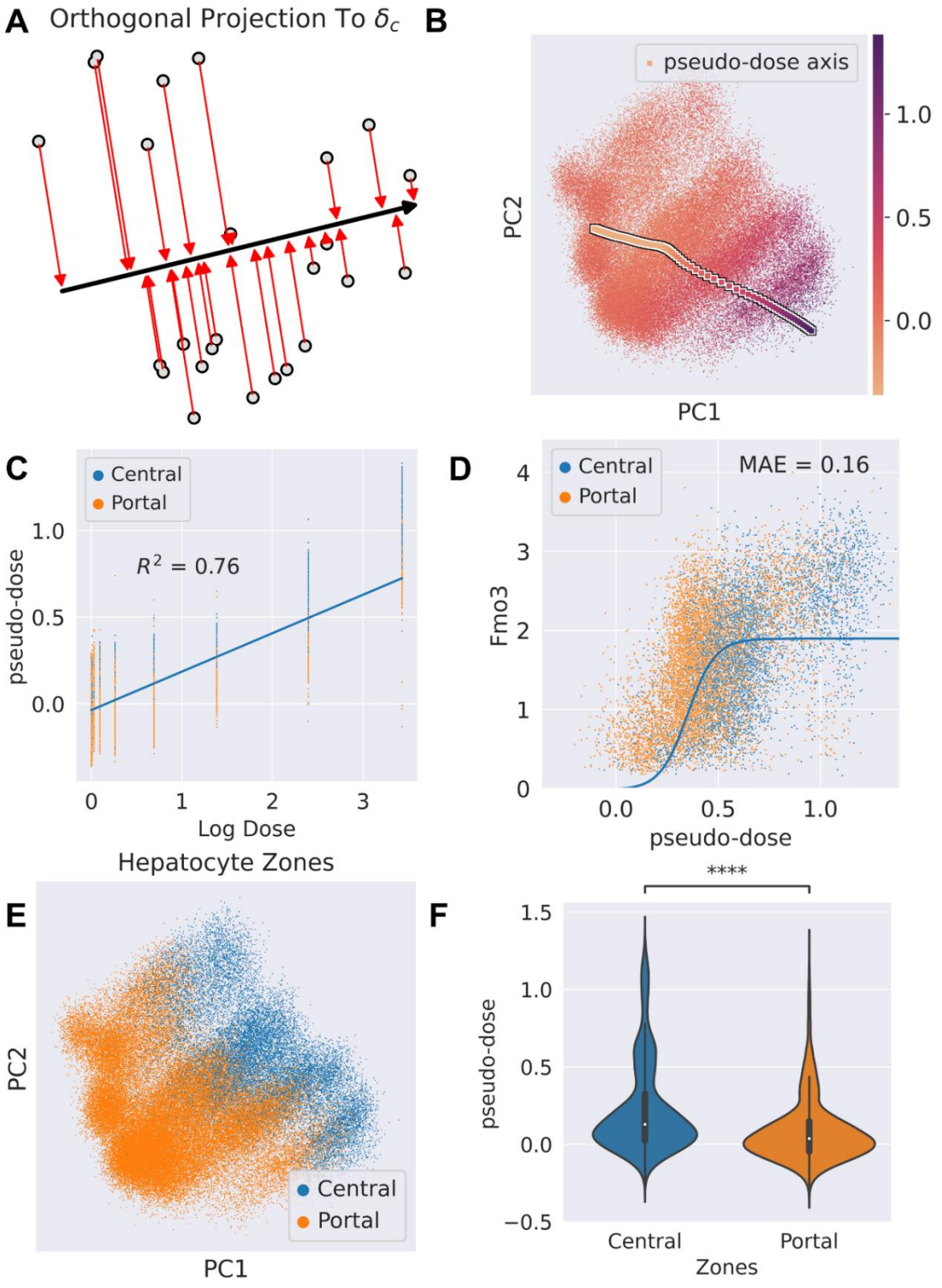

Figure 4.6 Interrogation of VAE using ridge regression in portal hepatocyte prediction. A) Schematic of calculating latent dimension weights using ridge regression. B) Bar plot of top 20 genes with the highest scVIDR genes scores. C) Enrichr analysis of the top 100 genes with

Figure 4.6 (cont'd)

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.

Working off the assumption that δ_c is the axis of perturbation in latent space, I orthogonally project the latent representation of each cell to the $span(\delta_c)$ to obtain a scalar coefficient for each cell along δ_c (Figure 4.6 A). I 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, I 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 central axis, but also their metabolic characteristics[53]. Thus, I 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 4.6C). I also found that pseudo-dose displayed a sigmoidal relationship (Section 4.4.6) between the expression of differentially expressed genes such as Fmo3 (Figure 4.6D). Finally, I found the pseudo-dose to be statistically higher on average in the central hepatocytes versus portal hepatocytes (Figure 4.6E). This is consistent with liver biology as central hepatocytes respond more strongly to treatment due to TCDD sequestration[59], and higher AhR expression levels in the centrilobular zone[113].

4.3 Discussion

When profiling any chemical, the dose often makes the poison. Development of dose-response prediction models is non-trivial and significant as low-dose changes to expression do not necessarily predict high dose changes. This is important as the effects seen at therapeutic doses for drugs do not necessarily extrapolate to overdose situations. To illustrate this, in figure 4.7 I show that changes in the first half of the TCDD dose-response of mouse liver portal-hepatocytes (between 0 and 0.3 μg/kg) doesn't correlate to changes in the second half of the dose-response (between 0.3 and 30 μg/kg). This is reflective of the non-linear nature of the chemical responses and belies the needs for methods that extrapolate dose-response utilizing other sources of data. I

show that scVIDR can handle predicting multiple doses from the chemical dose-responses of other cell types. I first modeled the dose response of thirty-eight chemicals across two datasets. I then show how I can interpret the dose-response predictions in scVIDR. Finally, I utilize the dimensionality reduction capabilities of scVIDR to order cells based on their individual cellular response using scVIDR.

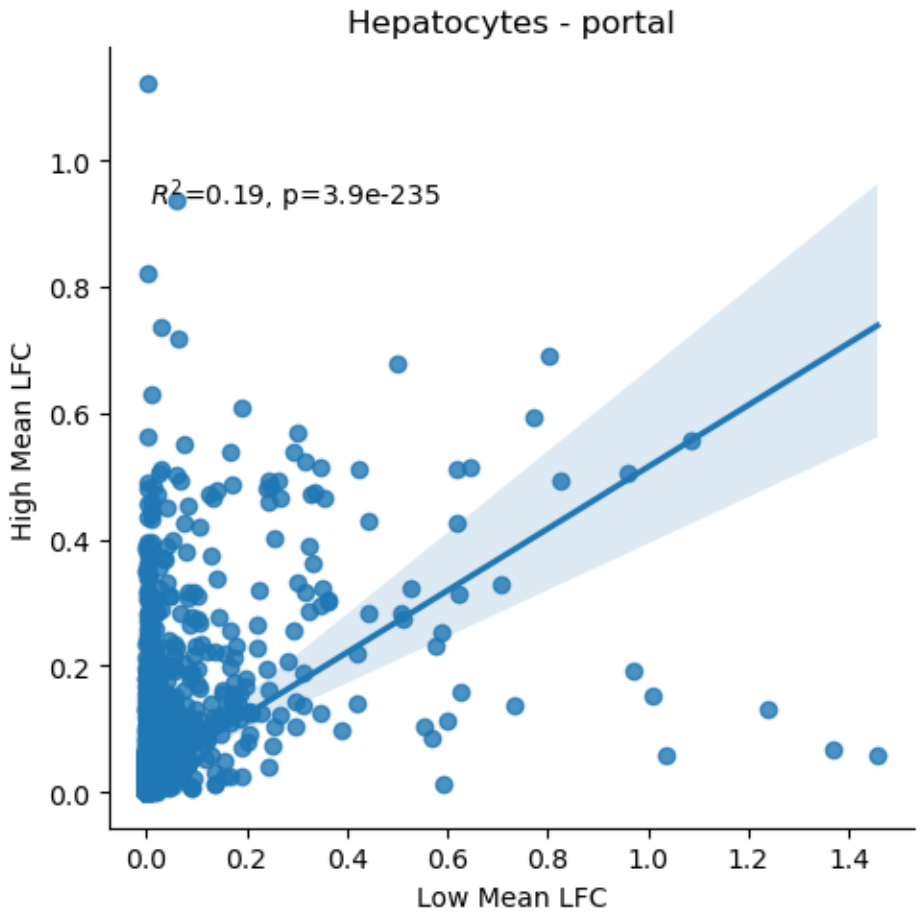

Figure 4.7 Regression plot of the magnitude of TCDD induced log fold-changes (LFC) of genes at low and high doses in mouse portal hepatocytes. All LFCs represent absolute values. Each point represents a gene. Low Mean LFC is the difference between the log(x+1) expression at dose 0 and 0.3 μg/kg. High Mean LFC is the difference between the log(x+1) expression at dose 0.3 and 30 μg/kg. The shaded region around the regression line represents the 95% confidence interval.

Examining the limitations of the model with respect to the Nault et al dataset[86], I can see that many of the same principles discussed in section 3.3 are relevant. Datasets with fewer control cells, and the less sensitive the cells, the worse the overall performance at higher doses (Figure 4.8). The notable exception is with stellate cells, where the highest dose is predicted worse in scVIDR than in scGen. The reason why is difficult to pin down completely. I hypothesize that this may be a result of the stellate cells starting to undergo an epithelial to mesenchymal transition to myofibroblasts[85]. This may be a result of the current formulation of the model not accounting for discrete transitions. In the future, it may be useful to use more sophisticated models to predict physiological transitions in cellular state like the epithelial to mesenchymal transition.

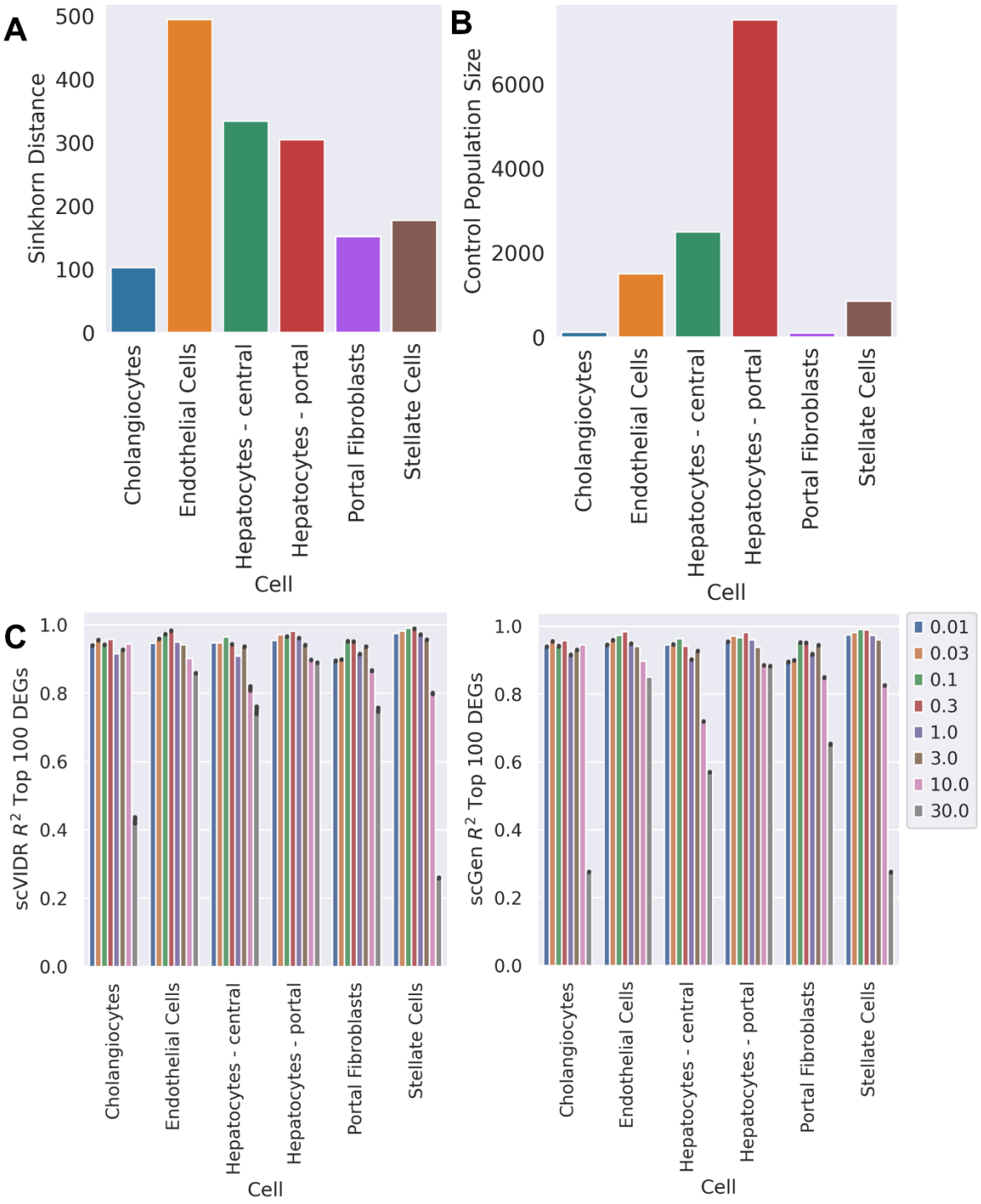

Figure 4.8 Impact of latent perturbation magnitude, and control population size on dose-response prediction performance. A) Sinkhorn distance between the latent distributions of the

Figure 4.8 (cont'd)

control and 30 μg/kg doses of TCDD of each cell type on the latent space in the Nault et al study[86]. C) Bar plot of the control group cell population size for each cell type. D) Bar plot of mean gene R^2 for each individual cell type when predicting all doses of TCDD in μg/kg.

In the sci-Plex dataset, prediction of certain drugs with epigenetic mode of actions produced the poorest prediction scores (Figure 4.9). This is because scSeq data provides no information regarding epigenetic modifications (e.g., chromatin accessibility, histone marks, and DNA binding proteins). Integration with epigenetic scSeq data such as single cell ATAC-seq could help to predict such responses with higher accuracy.

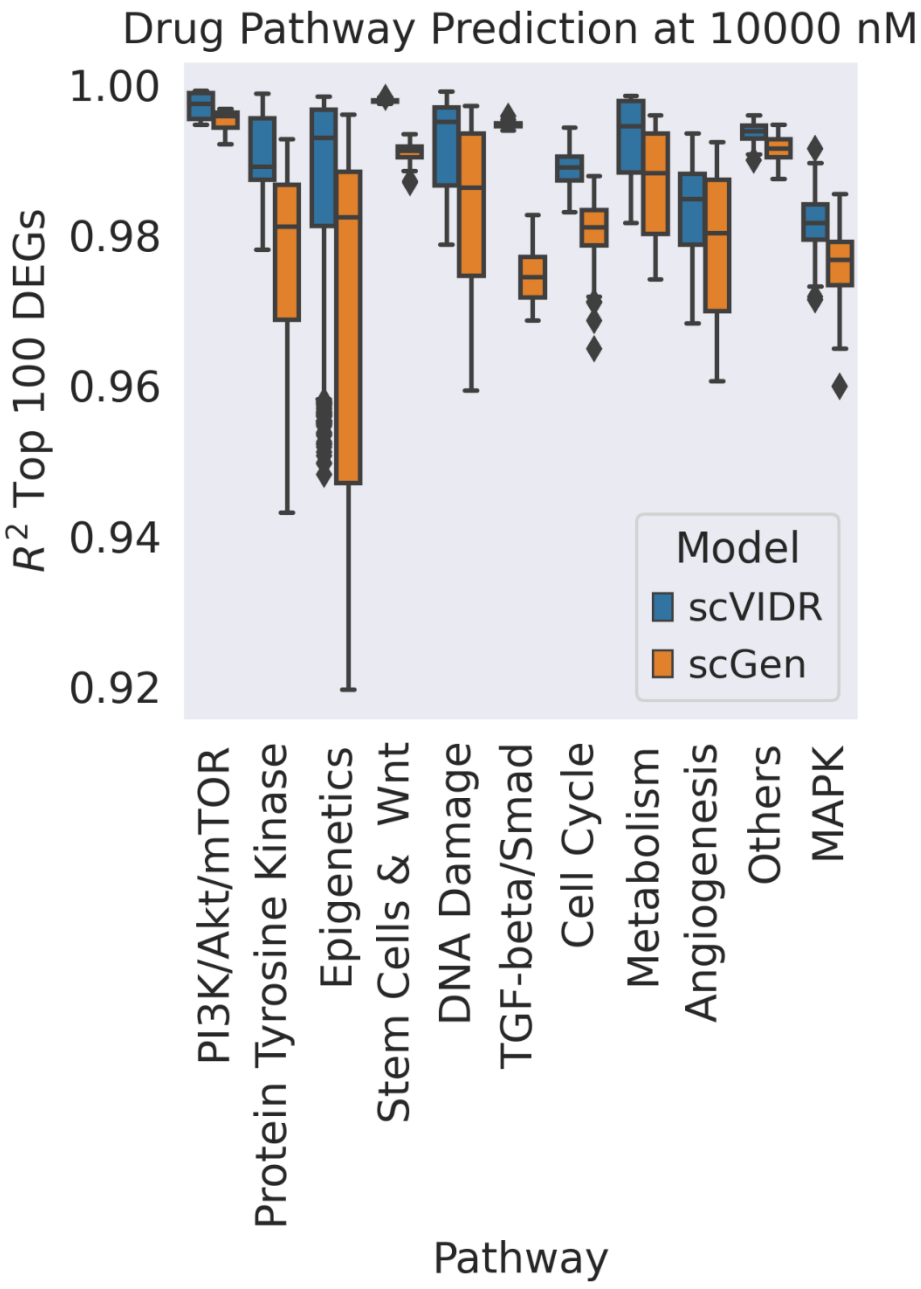

Figure 4.9 Overall drug pathway performances at the highest administered dose in Srivatsan et al. A boxplot of the mean gene R^2 Performance by scVIDR across all drug pathways in the test dataset at a dose of 10 μM.

While scVIDR and its pseudo-dose metric work on standard dose-response scenarios, it is likely inappropriate for use with more complex trajectories such as those found in cellular development and circadian rhythms[114]. Such trajectories include branching and cycling which involve

complex non-linear dynamics and require more sophisticated models to properly capture their topology. One could combine using scVIDR for dimensionality reduction and performing trajectory analysis using popular algorithms such as Monocle[112], slingshot[111], and diffusion pseudotime[56] to account for such trajectories.

Taken together, my 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. I 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 highest doses of previous experiments. As more data becomes 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.

4.4 Methods

4.4.1 Implementation of scVIDR

Implementation of scVIDR is identical to the implementation describe in sections 3.4.3 and 3.4.4.

4.4.2 Single cell expression datasets and preprocessing

The sci-Plex dataset[95] and TCDD dose-response dataset[86] 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, I select the top 5000 most highly variable genes to do my analysis on. The preprocessing was carried out using the *scanpy. pp* package using the *normalize_total, log1p,* and *highly_variable* functions[87].

The TCDD dose-response dataset comprised of single nuclei RNA-seq of C57BL6 of flash frozen mouse livers. Mice in this dataset were administered, sub-chronically, at a specified dose of TCDD via oral gavage every 4 days for 28 days. In my analysis, all immune cell types were left out, as immune cells are known to migrate from the lymph to the liver during TCDD administration[79]. Thus, there is a small size for the immune cell populations in the low-dose datasets versus the higher doses. The *sci-Plex* dataset is obtained from Srivatsan et al[95].

4.4.3 Interpolating on δ for dose-response prediction.

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

$$\hat{z}_{i,c,p=d} = z_{i,c,p=0} + p\delta$$

Where p is the coefficient, I want to calculate in order to calculate $\hat{z}_{i,c,p=d}$.

In log linear interpolation:

$$p = \frac{log(d+1)}{\log\left(max(d)+1\right)}$$

Where max(d) is the highest measured dose in the dataset.

In the threshold model:

$$p = \begin{cases} \frac{\log(d)}{\log(max(d))} & d > a \\ 0 & d \leq a \end{cases}$$

And in the Hill Model:

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

where a, L, k, d_0 are free parameters that I fit to the data. These free parameters were fit using Levenberg-Marquardt algorithm implementation in the *curve_fit* function in the *scipy.optimize* package.

4.4.4 Inferring feature level contributions to perturbation prediction

In PCA, I perform an orthogonal linear transformation on the data, such that my projected data preserves 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. I 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 genes and the principal components. I 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 size less than the observations) and a strictly linear map are nearly equivalent[115]. I can project principal components back into expression space using the following function:

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

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

$$\left\| X - \hat{X} \right\|_2^2$$

I find similarly that the loss function that I try to optimize in the autoencoder I described above is:

$$\| X - X W_1 W_2^T \|_2^2$$

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, I can see that the autoencoder described above can approximate the loadings of a principal component analysis 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 single layer with linear transfer operators such that the reconstruction error is the following:

$$\frac{1}{N} \| X - Z W_{Dec}^T \|^2$$

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. I can interpret W_{Dec} as a loadings matrix by which I can interpret the latent dimensions of the LDVAE.

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

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

Where Z are the sampled points from the latent space, ZW^T are the approximation of the predicted gene expression vectors, and W, is a m x n matrix where m Is the number of genes and n is the number of latent dimensions. I divide W using the $\left|\left|W\right|\right|_2$ to normalize for the effect of over expressed genes. I then calculate the gene scores by taking the dot product of normalized W and δ_c, or:

$$gene\ scores = \frac{W}{\left|\left|W\right|\right|_2} \cdot \delta_c$$

I use these gene scores to order genes for *Enrichr*[83] pathway analysis with the *gseapy* package[92]. Scores for each pathway were calculated using *score_genes* function from *scanpy.tl* package with the genes sets derived from the *Enrichr* results.

4.4.5 Calculating the pseudo-dose values.

I 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.

4.4.6 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 between the pseudo-dose and gene expression, a logistic function of the form:

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

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 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.

CHAPTER 5
CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Generative modeling for toxicological single cell RNA-seq

scRNA-seq has proliferated across many different disciplines of biology from development[116–118] to pharmacology[93,95]. In toxicology, single cell technologies have the potential to reveal the heterogeneous response previously masked by bulk tissue techniques like RNA-seq[17]. Even within cells of the same type there are exhibited differences in response depending on the transcriptional state of the cell[8,9]. While technologies like scRNA-seq have the potential to describe the entire spectrum of chemical response, due to the sheer volume of information and high-dimensionality, computational approaches must be deployed to uncover the biological complexity of tissue response.

Broadly, to deal with the high dimensionality of single cell data, bioinformaticists will utilize dimensionality reduction algorithms. Dimensionality reduction algorithms like PCA are applied routinely to do visualization and preprocessing for downstream analysis[36,37,42]. Over the last five years, VAEs (a deep generative model) have gained popularity as a dimensionality reduction algorithm in scRNA-seq analysis. This is due to the model's remarkable ability to perform data integration[29,34,48], and its ability to generate unseen scRNA-seq from its latent space[30,33].

In this dissertation I utilize variational autoencoders to perform batch correction for administration of TCDD to mice. First, I inferred the zonation of individual hepatocytes based on their gene expression. Then classified genes based on whether they are responsive to TCDD, exhibit a rhythmic expression profile, and/or exhibit a zonated expression profile. Due to the highly ordered nature of gene expression in the liver lobule, in previous studies[6,52,76] hepatocyte zonation is inferred from gene expression in control conditions. Unfortunately, TCDD significantly changes the expression of key zonation genes[62]. To accurately infer zonation from the liver, I performed batch correction in order to remove variance in expression caused by TCDD treatment. Using this new batch corrected dataset, I inferred zonation reducing the overall impact TCDD has on the inference. With the zonation properly inferred, I categorized hepatocyte genes based on their inferred zonation, time-series expression over twenty-four hours, and their response to TCDD treatment. With this gene classification, I revealed that most of the canonical TCDD response pathways exhibit both rhythmic and zonal properties and that these properties are independent of one another. This includes established spatial-temporal pathways in the liver

lobule including lipid and bile acid metabolism, and CYP450 metabolism pathways[52]. Additionally, I find that TCDD induced rhythmic changes are more associated with changes to the amplitude of the oscillations rather than changes in the phase. This follows with more than a third of TCDD affected genes completely losing rhythmicity. Mechanistically, I found that TCDD impacted the of the core circadian clock in the alternative *Arntl* binding partner to *Clock*, *Npas2*, and genes downstream of CLOCK-ARNTL binding such as *Per1*/2 and *Nr1d2*. Finally, I found that genes that had lost zonation were mostly associated with the central region of the hepatic lobule.

The generative properties of the VAE were utilized to predict chemical responses on unseen cell types in a model I developed called scVIDR. To do this, I built on the existing scGen framework and expanded it to predict across several doses. The main advantage of scVIDR over scGen is post-training regression on the latent space, which weights predictions of the response based on the cell's position on the latent space. This regression-based correction significantly improves prediction of chemical perturbations across several data sets such TCDD liver response, in-vitro cancer-responses to many different drug candidates, and PBMC response to IFN-β. This not only worked on unseen cell types, but I also utilized scVIDR to predict phagocyte response to LPS6 on unseen species. To predict multiple doses, I log-linearly interpolate on the latent space and predict the dose-response of TCDD in the liver and thirty-six cancer therapy drugs in A549 cells. The main advantage of such modeling is its ability to take existing single-cell chemical response datasets and utilize them to predict unseen chemical responses in novel cell types of interest.

Interpretation of the VAE latent space is difficult. In PCA, the lower dimensional principal component space can be interpreted using the weights of the linear functions compressing the data. These linear weights are like correlations between gene expression and the principal components. VAE's use non-linear neural networks to do their compression and thus the latent space of VAE's can't be easily interpreted. To interpret the latent space of the VAE, scVIDR approximates the decoder function using linear regression. The weights of the linear regression are used like the weights in PCA, which acts as an approximation of the relationship between the latent space and the dimensions of the of scRNA-seq measurement space. This allows the users to infer how changes in the position of latent space "correlate" to changes in gene expression measurements in scRNA-seq. Using this method, scVIDR predicts transcriptomic response of mouse hepatocytes to TCDD is most pronounced in canonical response genes like *Cyp1a1*.

Additionally, when performing gene pathway analysis on the most "correlated" genes, I identify canonical TCDD response pathways including CYP450 metabolism and fatty acid oxidation. Finally, responses to TCDD varies across individual cells, even those within the same cell type. I utilize scVIDR to measure such variation in the magnitude of response. To do this, I orthogonally project cells onto the δ_c, which is the mean difference between the control and treated cells on the latent space. The coefficient of the projection is the measure of the magnitude of response or the "pseudo-dose". With this pseudo-dose metric, I am able to demonstrate the difference in sensitivity between portal and central hepatocytes to TCDD.

5.2 Future directions in research the spatial-temporal organization of the liver lobule and associated pathologies

The liver is responsible for metabolism of nutrients and toxicants, production and recycling of proteins, and glucose and lipid homeostasis. Many of these tasks are zonated and rhythmic in the liver. Changes in the rhythmicity and zonation have been associated with NAFLD and drug induced liver injuries (DILI) such as those caused by acetaminophen. In this dissertation, I show that scRNA-seq can be used to comprehensively analyze responses of zonal and rhythmic processes to chemical perturbation of TCDD. I will first review future directions for studying TCDD's impact on zonation and rhythmicity. Then I will discuss alternative liver pathologies for which the methods of chapter 2 can be applied.

5.2.1 Future directions in studying TCDD's impact on rhythmicity and zonation and applications in other pathologies.

As described before, sub-chronic administration of TCDD leads to the ablation of rhythmicity in the mouse liver[66]. In acute administration, I see less severe effects on rhythmicity. Most effects on rhythmicity are seen in standard TCDD response pathways such as CYP450 metabolism. I see changes in the normal expression of the core circadian clock genes, agreeing with previous studies[66]. The most pronounced effect was increased expression of *Per2* across most of the acute response. Fader et al[66] has shown that CLOCK, ARNTL, and AhR colocalize on the *Per2* promoter, suggesting that TCDD induced AhR disruption of normal CLOCK-ARNTL binding may be the cause of this change. I also see higher expression *Per1* two hours post treatment followed by a decrease in expression for most time points after (with the exception of time point 18). Previous studies into the regulation of *Per1* have implicated β-naphthoflavone activated AhR in interrupting the normal binding of CLOCK-ARNTL in Hepa1c1c7 cells[84]. However, parsing how these acute changes in circadian rhythm leads to the

sub-chronic abolishment of rhythmicity in most core members of the circadian rhythm will require further studies on how different numbers of TCDD exposures change overall circadian clock expression and regulation.

TCDD exposure has also been shown to greatly alter zonation in sub-chronic exposures[62]. While impacts on zonation are predictably subtler within acute exposures, I confirm a bias towards pericentral expression seen in those same sub-chronic exposure studies[62]. Additionally, core regulatory pathways of zonation such as Wnt/β-catenin signaling were enriched in gene classes impacted by TCDD. How TCDD interrupts expression of these pathways is still unclear. TCDD induced changes to hepatocyte expression of Rspo3, which is a rheostat of zonation, may be the primary cause[62]. Additionally, I show that most changes to zonation are localized to the pericentral zone corroborating changes seen in pericentral *Apc*[62]. This pericentral localization of TCDD's effect on zonation is likely due to the high expression of AhR in the pericentral region[52] as well as activated Cyp1a2 sequestration of TCDD in the pericentral zone of the liver[119]. What is still unknown is how non-parenchymal response to TCDD is implicated in changes in zonation. Previous studies suggest that the rhythmic Wnt ligand expression of pericentral non-parenchymal cells is significant in zonation homeostasis[52]. Analysis on the potential impact TCDD has on this non-parenchymal signaling would potentially elucidate TCDD's impact on pericentral expression of zonation regulators.

One of the main limitations of the study regards inferring zonation from gene expression. Due to the nature of CYP450 expression being centrally enriched in liver lobules, TCDD biases gene expression towards more pericentral expression overall[62]. To get absolute localization of expression, spatial transcriptomic methodologies such as Xenium[120] or MERFISH[121] will be needed. Additionally, studies like those in chapter 2 are limited to cell types in the liver for which the zonation of expression is well profiled (i.e., hepatocytes[6,52] and endothelial cells[76,122]). Spatial experiments could better pinpoint the zonation patterns of genes across many cells in the non-parenchymal compartment of the liver such as cholangiocytes and immune cells. For example, resident macrophages, Kupffer cells, have been found to have an asymmetric distribution across the liver lobule, with more localizing in the periportal region[123]. However, their zonal expression profile has not been fully studied. Recent research in Kupffer cells have shown their transcription factor activity significantly altered with respect to TCDD exposure[124]. Additionally, in partial hepatectomy rodent models, Kupffer cells were important secretors of Wnt ligands in the

midlobular and periportal zones post loss of tissue[79,125]. This Wnt signaling goes on to propagate a loss of zonation very similar to the pericentral bias found in Nault et al[62] and in chapter 2. TCDD potentially could alter the expression of key Wnt ligand gene in Kupffer cells, much in the same way they do in partial hepatectomies[125]. The role of these and other non-parenchymal cells in TCDD liver zonation response, what affected hepatocytes they localize around, and what genes they differentially express proximal to liver injury can be better elucidated by spatial transcriptomic technologies. Additionally, insight generated by spatial experiments would have broader impacts in other zonated liver pathologies such as NAFLD and DILI.

5.2.2 Future Analysis of Rhythmicity and Zonation in Alternative Pathologies

Like TCDD administration in mice, many liver pathologies are characterized by perturbations to the spatial-temporal organization of the liver. More specifically, NAFLD has been connected to both zonal and temporal perturbations in liver metabolism[68]. Zonally, NAFLD is characterized by the accumulation of triglycerides due to increased influx of fatty acids in periportal hepatocytes and decreased efflux fatty acids in pericentral hepatocytes[68]. This results in steatosis localizing in pericentral hepatocytes[126]. Temporally, NAFLD has been observed in jet-lagged models of mice[127]. Additionally, molecular knock-out models of core-circadian genes show upticks in NAFLD and mortality[127]. Dissecting the entire etiology of NAFLD, will need more comprehensive descriptions of the perturbations to both zonation and temporal genes. Studies similar to those described in chapter 2 of this thesis could be utilized on NAFLD mouse models to identify these perturbations and better describe the progression of NAFLD pathologies.

Drug induced liver injuries (DILI) exhibit distinct zonal properties. In carbon tetrachloride (CCL4) models of DILI (which cause peri-central injury), injury leads to regenerative responses from adjacent layers of hepatocytes such as those found in the periportal region[128]. These regenerative responses were accompanied by a temporary loss of zonation which was recovered six days post DILI[129]. It is likely this loss of zonation is connected to the activation of proliferative phenotypes post DILI[128]. Identifying which genes lose their zonation in response to DILI, and where proliferative genes are being turned on along the zonation axis of the liver lobule will lead to stronger understandings of liver regeneration and DILI prevention.

DILI is also dependent on circadian rhythm. For example, alanine transferase (ALT) levels were higher in mice administered with acetaminophen at ZT14 vs. ZT2[130]. This is correlated with peaks in acetaminophen enzyme expression early in the night. However, these effects are

dependent on feeding patterns as fasting induced higher ALT levels when compared to control. Despite this, CLOCK-deficient mice had resilient ALT levels regardless of acetaminophen administration time. Thus, DILI can be dependent on the circadian clock of the animal[130]. Identifying genes in a drug's mode of action that are rhythmic would be helpful in determining more optimal times in which to administer said drug as to prevent DILIs.

5.3 Future Directions in Variational Autoencoders and the Prediction of Chemical Perturbations

Illustrated already in this dissertation is scVIDR's ability to improve on perturbation prediction and predict additional doses along the dose-response. Next steps include using these models to infer important dose-response metrics such as potency and maximal efficacy. Additionally, relating the trajectory of the pseudo-dose inferred by scVIDR to potential adverse outcome pathways or using it to infer new ones would be incredibly useful for toxicologists. Finally, I will discuss potential improvements to scVIDR with these end goals in mind.

5.3.1 Broader Toxicological Impacts of scVIDR

One of the most prominent features a toxicologist wants to extract is the drug's potency. Several dose-response metrics measure potency. For example, EC50 is the concentration at which the drug elicits fifty percent of its maximal response (assuming a standard dose-response)[131]. In section 4.2.2 a hill-model of interpolation is discussed for prediction but is ultimately discarded due to a lack of significant improvement over the more parsimonious log-linear model. However, models like this would, in theory, predict the EC50 (d_0 parameter from section 4.4.3). Validation of these predictions would require an external dataset of EC50s specific to cell types and thus was not included in the main text. Ensuring accurate predictions for chemical potency with models like scVIDR will rely on improving datasets as well as improving on the current model formulation. Current proposals for single cell perturbational atlases[132] as well as the continuation of initiatives like TOXCAST21[133] stand to improve on this issue.

Ultimately, in addition to parameter discovery, toxicologists would like to use models like scVIDR to predict adverse outcome pathways. Adverse outcome pathways (AOPs) are the sequence of cellular and molecular events that lead a chemical to elicit a toxic effect[134]. I can relate δ or the "pseudo-dose" to gene pathways using existing methods such as gene set enrichment analysis (GSEA) as described in section 4.2.4. Expanding on the pseudo-dose algorithm to account for branching trajectories will better account for distinguishing between different sources of toxicity. One major limitation of scVIDR currently is that it can only predict

transcriptomic changes associated with a chemical. How well transcriptional changes reflect changes in protein levels and protein activity is a matter of intense debate[135,136]. While within sample correlations are rather robust, across samples (i.e., across tissues or conditions) correlations are much more modest due to variances in post-transcriptional and translational kinetics[136]. So, while increases or decrease in RNA expression likely imply increases or decreases in protein expression, the magnitude of these changes is much less correlated due to post-transcriptional, translational, and post-translational mechanisms of protein regulation. This protein regulation is also a frequent feature of AOPs[137]. Thus, understanding the toxic response from only the transcriptomic perspective would immediately limit the scope of discovery. Thus, integration of other data modalities such as CITE-seq and scATAC-seq is required address this limitation.

5.3.2 Improving scVIDR.

The main determinant of scVIDR's prediction accuracy is in calculation of the δ. This is especially important when wanting to improve predictions of cell type specific EC50s and AOPs. While I show marked improvements in this prediction, potentially the model can be further improved with a change to the resolution of the prediction. Here, instead of cell-type specific δ_c, I calculate a δ_i, where i is the index of a specific cell. This would allow for a more fine-grained prediction of cellular response. The main limitation in this approach would be how to calculate individual δ_i as individual cells cannot be followed across the dose-response using scRNA-seq. This is because scRNA-seq destroys the cells during sequencing. Alternative technologies such as Live-seq[138] would be a way to overcome this limitation. However, no commercial platform yet exists with this type of technology for wide-spread adoption. Alternatively, a computational solution would be to map the control cells to the treatment cells on the latent space. The mapping would be used to infer where on the latent treatment distribution a control cell would lie after perturbation, thus giving us point pairs to calculate δ_i. Having an individualized δ_i would allow for more complicated models, such as neural networks, to predict δ_i for unseen cells. For example, scANVI[46] utilizes a neural network, $q(c_i|z_i)$, on the latent space to classify cells into their respective cell types. This framework could be adapted to predict δ_i for unseen cells, by taking the existing δ_i on the latent space and using them as training for the neural network, $q(\delta_i|z_i)$. However, this assumes that the shortest path on the latent space is the best representation of the underlying biological manifold. Optimal transport-based methods such as

TrajectoryNet[139] utilize continuous normalizing flows[140] to infer the underlying trajectory of cells along some time-series experiment. In theory, one could adapt this to the trajectory of the dose-response to get more accurate uncertainties in predictions. Future work will be needed to study how to improve prediction accuracy for concrete toxicological endpoints and drug safety studies. The pseudo-dose metric works for standard dose-response scenarios described in this dissertation. However, it struggles when faced with more complex trajectories. These more complex trajectories could be observed in relation to chrono-pharmacology[66] or chemical disruptions to cellular development[141]. Additionally, toxicological responses can lead to more than one type of cell injury. To account for this, one could extend pseudo-dose using methods such as TSCAN[142] or diffusion pseudo-time[56]. The only change the user would be to preprocess using scVIDR instead of PCA or UMAP. Additionally, with more complex trajectories, toxicologists would like to be able to distinguish between different sources of response (i.e., cell death vs. cellular necrosis). Differential trajectory analysis methods like condiments[143] or Trade-seq[21] could be used to identify different genes or different sets of genes associated with different branches or points in the dose-response. These genes could be used to identify potential mechanisms by which the drug will cause response.

Currently, scVIDR doesn't integrate other data modalities important to toxic response such as chromatin accessibility[144] and protein expression[145]. Already, scVIDR has shown that it sometimes performs poorly on drugs with epigenetic modes of action (section 4.3; Figure 4.8). Incorporation of other data modalities will be critical not only for future models of chemical response prediction, but also in describing AOPs associated with different chemicals. VAE models that aim to integrate different types of single cell technologies such as MultiVI present a potential path forward. For example, let us assume that I have scRNA-seq, scATAC-seq, and CITE-seq datasets with some overlapping cell-types. Using an integrative version of scVIDR, multiple single cell technologies would be integrated to a single latent space. Thus, similar predictions could be extended to changes in chromatin accessibility with the scATAC-seq, and changes in expression of cell surface proteins with CITE-seq. With this integration, scVIDR will not only be able to predict transcriptomic events in an adverse outcome pathway, but also epigenetic and cell surface protein events.

5.4 Closing remarks

In this dissertation I utilize current and develop new frameworks in generative modeling for toxicology. I show VAEs utility to describe responses to chemicals in different scRNA-seq datasets. My discussions in rhythmicity and zonation in TCDD response have implications in other liver pathologies such as NAFLD and DILI. Time-series modeling with VAEs could be incorporated similarly to help with these types of experiments. While VAEs are incredibly useful for scRNA-seq, high data requirements for training (i.e., thousands of cells) mean they require thousands of bulk RNA-seq expression datasets be properly trained on existing resources such as TOXCAST21[133]. Despite this, VAEs show great promise in aiding toxicologists in studying transcriptomic response. With larger single cell initiatives, I envision models like scVIDR can be extended to predict EC50s and adverse outcome pathways. Additionally, VAEs ability to integrate different types of single cell data, such as scATAC-seq and CITE-seq, will allow for better prediction of chemical response. With the proliferation of single cell technologies, an opportunity to understand toxicology at cellular resolution has presented itself. However, high dimensionality makes analysis challenging. I have shown that variational autoencoders can be used to process such data to make insights, as well as generate novel predictions using existing toxicological data.

BIBLIOGRAPHY

1. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 16, 3–50 (1996).
2. Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science (1979)* **274**, 1531–1534 (1996).
3. Sun, D., Gao, W., Hu, H. & Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B* **12**, 3049–3062 (2022).
4. Aissa, A. F. *et al.* Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat Commun* **12**, 1628 (2021).
5. null, null *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science (1979)* **376**, eabl4896 (2023).
6. Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature 2017 542:7641* **542**, 352–356 (2017).
7. Raj, A. & van Oudenaarden, A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* vol. 135 216–226 Preprint at https://doi.org/10.1016/j.cell.2008.09.050 (2008).
8. Yao, J., Pilko, A. & Wollman, R. Distinct cellular states determine calcium signaling response. *Mol Syst Biol* **12**, 894 (2016).
9. Kramer, B. A. & Pelkmans, L. Cellular state determines the multimodal signaling response of single cells. *bioRxiv* 2019.12.18.880930 (2019).
10. Yang, Y., Filipovic, D. & Bhattacharya, S. 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* (2021) doi:10.1002/hep4.1848.
11. Madeira, C. & Costa, P. M. Chapter Two - Proteomics in systems toxicology. in *Proteomics and Systems Biology* (eds. Donev, R. & Karabencheva-Christova, T. B. T.-A. in P. C. and S. B.) vol. 127 55–91 (Academic Press, 2021).
12. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–35 (2006).
13. Ma, R., Martínez-Ramírez, A. S., Borders, T. L., Gao, F. & Sosa-Pineda, B. Metabolic and non-metabolic liver zonation is established non-synchronously and requires sinusoidal Wnts. *Elife* **9**, (2020).
14. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
15. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

16. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 1–12 (2017).

17. Zhang, Q. *et al.* Embracing systems toxicology at single-cell resolution. *Curr Opin Toxicol* **16**, 49–57 (2019).

18. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901).

19. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

20. Ouyang, J. F., Kamaraj, U. S., Cao, E. Y. & Rackham, O. J. L. ShinyCell: simple and sharable visualization of single-cell gene expression data. *Bioinformatics* **37**, 3374–3376 (2021).

21. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun* **11**, 1201 (2020).

22. Linderman, G. C. Dimensionality Reduction of Single-Cell RNA-Seq Data. in *RNA Bioinformatics* (ed. Picardi, E.) 331–342 (Springer US, New York, NY, 2021). doi:10.1007/978-1-0716-1307-8_18.

23. Fefferman, C., Mitter, S. & Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society* **29**, 983–1049 (2016).

24. Whitacre, J. M. Degeneracy: a link between evolvability, robustness and complexity in biological systems. *Theor Biol Med Model* **7**, 6 (2010).

25. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. Evolution of genetic redundancy. *Nature* **388**, 167–171 (1997).

26. Chen, B. *et al.* Relating Chemical Structure to Cellular Response: An Integrative Analysis of Gene Expression, Bioactivity, and Structural Data Across 11,000 Compounds. *CPT Pharmacometrics Syst Pharmacol* **4**, 576–584 (2015).

27. Davidson, E. H. The “Regulatory Genome” for Animal Development. in *The Regulatory Genome* 1–29 (Elsevier, 2006). doi:10.1016/b978-012088563-3.50019-5.

28. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2014).

29. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053–1058 (2018).

30. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat Methods* **16**, 715–721 (2019).

31. Grønbech, C. H. *et al.* scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics* **36**, 4415–4422 (2020).

32. Qiu, Y. L., Zheng, H. & Gevaert, O. Genomic data imputation with variational auto-encoders. *Gigascience* **9**, 1–12 (2020).

33. Lotfollahi, M., Naghipourfar, M., Theis, F. J. & Wolf, F. A. Conditional out-of-distribution generation for unpaired data using transfer VAE. *Bioinformatics* **36**, i610–i617 (2020).

34. Ashuach, T. *et al.* MultiVI: deep generative model for the integration of multimodal data. *Nat Methods* **20**, 1222–1231 (2023).

35. Bank, D., Koenigstein, N. & Giryes, R. Autoencoders. in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (eds. Rokach, L., Maimon, O. & Shmueli, E.) 353–374 (Springer International Publishing, Cham, 2023). doi:10.1007/978-3-031-24628-9_16.

36. Liu, Z. Visualizing single-cell RNA-seq data with Semisupervised principal component analysis. *Int J Mol Sci* **21**, 5797 (2020).

37. Grabski, I. N., Street, K. & Irizarry, R. A. Significance analysis for clustering with single-cell RNA-sequencing data. *Nat Methods* **20**, 1196–1202 (2023).

38. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).

39. Marini, F. & Binder, H. pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics* **20**, 331 (2019).

40. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLoS Comput Biol* **19**, e1011288- (2023).

41. Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Mol Syst Biol* **19**, e11517 (2023).

42. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**, 550–572 (2023).

43. Higgins, I. *et al.* Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579* (2016).

44. Ding, J., Condon, A. & Shah, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat Commun* **9**, (2018).

45. Wang, D. & Gu, J. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genomics Proteomics Bioinformatics* **16**, 320–331 (2018).

46. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol* **17**, e9620 (2021).

47. Jiang, J. *et al.* Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder. *Brief Bioinform* **24**, bbad152 (2023).

48. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* **40**, 121–130 (2022).

49. Liu, S., Zhang, Y., Peng, J. & Shang, X. An improved hierarchical variational autoencoder for cell–cell communication estimation using single-cell RNA-seq data. *Brief Funct Genomics* elac056 (2023) doi:10.1093/bfgp/elac056.

50. Yao, Z. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023).

51. Kopf, A., Fortuin, V., Somnath, V. R. & Claassen, M. Mixture-of-Experts Variational Autoencoder for clustering and generating from similarity-based representations on single cell data. *PLoS Comput Biol* **17**, e1009086 (2021).

52. Droin, C. *et al.* Space-time logic of liver gene expression at sub-lobular scale. *Nat Metab* **3**, 43–58 (2021).

53. Cunningham, R. P. & Porat-Shliom, N. Liver Zonation – Revisiting Old Questions With New Technologies. *Front Physiol* **12**, 732929 (2021).

54. Dibner, C., Schibler, U. & Albrecht, U. The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks. *Annu Rev Physiol* **72**, 517–549 (2010).

55. Buijs, R. M. *et al.* Organization of circadian functions: interaction with the body. in *Progress in Brain Research* (eds. Kalsbeek, A. et al.) vol. 153 341–360 (Elsevier, 2006).

56. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845–848 (2016).

57. Coifman, R. R. & Lafon, S. Diffusion maps. *Appl Comput Harmon Anal* **21**, 5–30 (2006).

58. Plumlee, K. Chapter 11. Hepatobiliary System. in 61–68 (2004). doi:10.1016/B0-32-301125-X/50014-5.

59. Santostefano, M. J. *et al.* Dose-dependent localization of TCDD in isolated centrilobular and periportal hepatocytes. *Toxicological Sciences* **52**, 9–19 (1999).

60. Jungermann, K. Zonation of metabolism and gene expression in liver. *Histochem Cell Biol* **103**, 81–91 (1995).

61. Wild, S. L. *et al.* The canonical Wnt pathway as a key regulator in liver development, differentiation and homeostatic renewal. *Genes (Basel)* **11**, 1163 (2020).

62. Nault, R. *et al.* Single-cell transcriptomics shows dose-dependent disruption of hepatic zonation by TCDD in mice. *Toxicological Sciences* **191**, 135–148 (2023).

63. Pai, S. G. *et al.* Wnt/beta-catenin pathway: modulating anticancer immune response. *J Hematol Oncol* **10**, 101 (2017).

64. Green, C. B., Takahashi, J. S. & Bass, J. The meter of metabolism. *Cell* **134**, 728–742 (2008).

65. Yeung, J. & Naef, F. Rhythms of the Genome: Circadian Dynamics from Chromatin Topology, Tissue-Specific Gene Expression, to Behavior. *Trends in Genetics* **34**, 915–926 (2018).

66. Fader, K. A., Nault, R., Doskey, C. M., Fling, R. R. & Zacharewski, T. R. 2,3,7,8-Tetrachlorodibenzo-p-dioxin abolishes circadian regulation of hepatic metabolic activity in mice. *Sci Rep* **9**, 6514 (2019).

67. Shi, D. *et al.* Circadian clock genes in the metabolism of non-alcoholic fatty liver disease. *Front Physiol* **10**, 423 (2019).

68. Martini, T., Naef, F. & Tchorz, J. S. Spatiotemporal Metabolic Liver Zonation and Consequences on Pathophysiology. *Annual Review of Pathology: Mechanisms of Disease* **18**, 439–466 (2023).

69. Tahara, Y. & Shibata, S. Circadian rhythms of liver physiology and disease: experimental and clinical evidence. *Nat Rev Gastroenterol Hepatol* **13**, 217–226 (2016).

70. Procházková, J. *et al.* The interplay of the aryl hydrocarbon receptor and β-catenin alters both AhR-dependent transcription and wnt/β-catenin signaling in liver progenitors. *Toxicological Sciences* **122**, 349–360 (2011).

71. Schneider, A. J., Branam, A. M. & Peterson, R. E. Intersection of AHR and Wnt signaling in development, health, and disease. *International Journal of Molecular Sciences* vol. 15 17852–17885 Preprint at https://doi.org/10.3390/ijms151017852 (2014).

72. Time-Course Single-Nuclei RNA-Seq Analysis Identifies NRG/ERBB Signaling Dysregulation in TCDD-Induced Hepatotoxicity. *Society of Toxicology Annual Meeting* Preprint at (2023).

73. Larigot, L., Juricek, L., Dairou, J. & Coumoul, X. AhR signaling pathways and regulatory functions. *Biochim Open* **7**, 1–9 (2018).

74. Hutin, D. *et al.* 2,3,7,8-Tetrachlorodibenzo-p-Dioxin (TCDD)-Inducible Poly-ADP-Ribose Polymerase (TIPARP/PARP7) Catalytic Mutant Mice (TiparpH532A) Exhibit Increased Sensitivity to TCDD-Induced Hepatotoxicity and Lethality. *Toxicological Sciences* **183**, 154–169 (2021).

75. Davidian, M. & Giltinan, D. M. Nonlinear models for repeated measurement data: An overview and update. *J Agric Biol Environ Stat* **8**, 387 (2003).

76. Aizarani, N. *et al.* A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).

77. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022).

78. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* **17**, 75 (2016).

79. Nault, R., Fader, K. A., Bhattacharya, S. & Zacharewski, T. R. Single-Nuclei RNA Sequencing Assessment of the Hepatic Effects of 2,3,7,8-Tetrachlorodibenzo-p-dioxin. *CMGH* **11**, 147–159 (2021).

80. Schmidt, C. K. *et al.* 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) alters the endogenous metabolism of all-trans-retinoic acid in the rat. *Arch Toxicol* **77**, 371–383 (2003).

81. Burke, Z. D. & Tosh, D. The Wnt/β-catenin pathway: master regulator of liver zonation? *BioEssays* **28**, 1072–1077 (2006).

82. Gougelet, A. *et al.* T-cell factor 4 and β-catenin chromatin occupancies pattern zonal liver metabolism in mice. *Hepatology* **59**, 2344–2357 (2014).

83. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

84. Xu, C.-X., Krager, S. L., Liao, D.-F. & Tischkau, S. A. Disruption of CLOCK-BMAL1 Transcriptional Activity Is Responsible for Aryl Hydrocarbon Receptor–Mediated Regulation of Period1 Gene. *Toxicological Sciences* **115**, 98–108 (2010).

85. Nault, R. *et al.* Dose-Dependent Metabolic Reprogramming and Differential Gene Expression in TCDD-Elicited Hepatic Fibrosis. *Toxicological Sciences* **154**, 253–266 (2016).

86. Nault, R. *et al.* Benchmarking of a Bayesian single cell RNAseq differential gene expression test for dose–response study designs. *Nucleic Acids Res* (2022) doi:10.1093/nar/gkac019.

87. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).

88. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* vol. 57 10–25080 (Austin, TX, 2010).

89. Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *Comput J* **7**, 308–313 (1965).

90. Schwarz, G. Estimating the dimension of a model. *The annals of statistics* 461–464 (1978).

91. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods 2020 17:3* **17**, 261–272 (2020).

92. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).

93. Peidli, S. *et al.* scPerturb: Harmonized Single-Cell Perturbation Data. *bioRxiv* 2022.08.20.504663 (2023) doi:10.1101/2022.08.20.504663.

94. McFarland, J. M. *et al.* Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat Commun* **11**, (2020).

95. Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution. *Science (1979)* **367**, 45–51 (2020).

96. Ji, Y., Lotfollahi, M., Wolf, F. A. & Theis, F. J. Machine learning for perturbational single-cell omics. *Cell Syst* **12**, 522–537 (2021).

97. Bunne, C. *et al.* Learning Single-Cell Perturbation Responses using Neural Optimal Transport. *bioRxiv* 2021.12.15.472775 (2021) doi:10.1101/2021.12.15.472775.

98. Wei, X., Dong, J. & Wang, F. scPreGAN, a deep generative model for predicting the response of single-cell expression to perturbation. *Bioinformatics* **38**, 3377–3384 (2022).

99. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89–94 (2018).

100. Hagai, T. *et al.* Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197–202 (2018).

101. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825–2830 (2011).

102. Arpit, D., Zhou, Y., Ngo, H. & Govindaraju, V. Why regularized auto-encoders learn sparse representation? in *International Conference on Machine Learning* 136–144 (PMLR, 2016).

103. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science (1979)* **297**, 1183–1186 (2002).

104. Rostom, R., Svensson, V., Teichmann, S. A. & Kar, G. Computational approaches for interpreting scRNA-seq data. *FEBS Lett* **591**, 2213–2225 (2017).

105. Ribeiro, M. T., Singh, S. & Guestrin, C. ' Why should i trust you?' Explaining the predictions of any classifier. in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* 1135–1144 (2016).

106. Uno, S. *et al.* Cyp1a1(−/−) male mice: protection against high-dose TCDD-induced lethality and wasting syndrome, and resistance to intrahepatocyte lipid accumulation and uroporphyria. *Toxicol Appl Pharmacol* **196**, 410–421 (2004).

107. Martens, M. *et al.* WikiPathways: Connecting communities. *Nucleic Acids Res* **49**, D613–D621 (2021).

108. Henry, E. C., Welle, S. L. & Gasiewicz, T. A. TCDD and a Putative Endogenous AhR Ligand, ITE, Elicit the Same Immediate Changes in Gene Expression in Mouse Lung Fibroblasts. *Toxicological Sciences* **114**, 90–100 (2010).

109. Cholico, G. N. *et al.* Thioesterase induction by 2,3,7,8-tetrachlorodibenzo-p-dioxin results in a futile cycle that inhibits hepatic β-oxidation. *Sci Rep* **11**, 15689 (2021).

110. Friedrich, M. *et al.* Tryptophan metabolism drives dynamic immunosuppressive myeloid states in IDH-mutant gliomas. *Nat Cancer* **2**, 723–740 (2021).

111. Street, K. *et al.* Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

112. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982 (2017).

113. Lindros, K. O., Oinonen, T., Johansson, I. & Ingelman-Sundberg, M. Selective Centrilobular Expression of the Aryl Hydrocarbon Receptor in Rat Liver. *Journal of Pharmacology and Experimental Therapeutics* **280**, 506 LP – 511 (1997).

114. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547–554 (2019).

115. Plaut, E. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253* (2018).

116. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

117. Byrnes, L. E. *et al.* Lineage dynamics of murine pancreatic development at single-cell resolution. *Nat Commun* **9**, 3922 (2018).

118. Pellin, D. *et al.* A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat Commun* **10**, (2019).

119. Diliberto, J. J., Burgin, D. & Birnbaum, L. S. Role of CYP1A2 in Hepatic Sequestration of Dioxin: Studies Using CYP1A2 Knock-Out Mice. *Biochem Biophys Res Commun* **236**, 431–433 (1997).

120. Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun* **14**, 8353 (2023).

121. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (1979)* **348**, aaa6090 (2015).

122. Halpern, K. B. *et al.* Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat Biotechnol* **36**, 962–970 (2018).

123. Gola, A. *et al.* Commensal-driven immune zonation of the liver promotes host defence. *Nature* **589**, 131–136 (2021).

124. Karri, K. & Waxman, D. J. TCDD dysregulation of lncRNA expression, liver zonation and intercellular communication across the liver lobule. *Toxicol Appl Pharmacol* **471**, 116550 (2023).

125. Walesky, C. M. *et al.* Functional compensation precedes recovery of tissue mass following acute liver injury. *Nat Commun* **11**, 5785 (2020).

126. Brunt, E. M. Pathology of nonalcoholic fatty liver disease. *Nat Rev Gastroenterol Hepatol* **7**, 195–203 (2010).

127. Kettner, N. M. *et al.* Circadian homeostasis of liver metabolism suppresses hepatocarcinogenesis. *Cancer Cell* **30**, 909–924 (2016).

128. Panday Chase P.; Khetani Salman R., R. M. The Role of Liver Zonation in Physiology, Regeneration, and Disease. *Semin Liver Dis* **42**, 1–16 (2022).

129. Wei, Y. *et al.* Liver homeostasis is maintained by midlobular zone 2 hepatocytes. *Science (1979)* **371**, eabb1625 (2021).

130. DeBruyne, J. P., Weaver, D. R. & Dallmann, R. The Hepatic Circadian Clock Modulates Xenobiotic Metabolism in Mice. *J Biol Rhythms* **29**, 277–287 (2014).

131. Jiang, X. & Kopp-Schneider, A. Summarizing EC50 estimates from multiple dose-response experiments: A comparison of a meta-analysis strategy to a mixed-effects model approach. *Biometrical Journal* **56**, 493–512 (2014).

132. Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Mol Syst Biol* **19**, e11517 (2023).

133. R, T. R., P, A. C., J, K. R. & R, B. J. Improving the Human Hazard Characterization of Chemicals: A Tox21 Update. *Environ Health Perspect* **121**, 756–765 (2013).

134. Sakuratani, Y., Horie, M. & Leinala, E. Integrated Approaches to Testing and Assessment: OECD Activities on the Development and Use of Adverse Outcome Pathways and Case Studies. *Basic Clin Pharmacol Toxicol* **123**, 20–28 (2018).

135. Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J. & Smith, V. A. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci Rep* **5**, 10775 (2015).

136. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**, 630–644 (2020).

137. Liu, A., Han, N., Munoz-Muriedas, J. & Bender, A. Deriving time-concordant event cascades from gene expression data: A case study for Drug-Induced Liver Injury (DILI). *PLoS Comput Biol* **18**, e1010148- (2022).

138. Chen, W. *et al.* Live-seq enables temporal transcriptomic recording of single cells. *Nature* **608**, 733–740 (2022).

139. Tong, A., Huang, J., Wolf, G., Van Dijk, D. & Krishnaswamy, S. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. in *International conference on machine learning* 9526–9536 (PMLR, 2020).

140. Mathieu, E. & Nickel, M. Riemannian continuous normalizing flows. *Adv Neural Inf Process Syst* **33**, 2503–2515 (2020).

141. Khan, D. M. I. O., Karmaus, P. W. F., Bach, A., Crawford, R. B. & Kaminski, N. E. An in vitro model of human hematopoiesis identifies a regulatory role for the aryl hydrocarbon receptor. *Blood Adv* **7**, 6253–6265 (2023).

142. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* **44**, e117–e117 (2016).

143. Bézieux, H. R. de, Berge, K. Van den, Street, K. & Dudoit, S. Trajectory inference across multiple conditions with condiments: differential topology, progression, differentiation, and expression. *bioRxiv* 2021.03.09.433671 (2021) doi:10.1101/2021.03.09.433671.

144. Davis, I. J. & Pattenden, S. G. Chapter 1-3 - Chromatin Accessibility as a Strategy to Detect Changes Associated With Development, Disease, and Exposure and

Susceptibility to Chemical Toxins. in *Toxicoepigenetics* (eds. McCullough, S. D. & Dolinoy, D. C.) 85–103 (Academic Press, 2019). doi:https://doi.org/10.1016/B978-0-12-812433-8.00003-4.

145. Kennedy, S. The role of proteomics in toxicology: identification of biomarkers of toxicity by protein expression analysis. *Biomarkers* **7**, 269–290 (2002).