

MODELING THE SPATIO -TEMPORAL EFFECT OF 2,3,7,8 TETRACHLORODIBENZO-P-
DIOXIN ON HEPATIC GENE EXPRESSION

By

Daniel Kwabena Marri

A DISSERTATION

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

Biomedical Engineering – Doctor of Philosophy

2024

ABSTRACT

Circadian clocks are intrinsic molecular oscillators present in cells across prokaryotes and eukaryotes that synchronize physiological processes with external cues, enabling organismal adaptation and survival. These clocks regulate crucial biological functions, including sleep-wake cycles, thermoregulation, hepatic metabolism, and hormonal secretion, through the rhythmic expression of clock-controlled genes. The mammalian liver comprises structural units called lobules, with hepatocytes arranged in a hexagonal pattern along a pericentral-to-periportal axis extending from the central vein to the portal triad. Perturbations in the circadian clock network can contribute to the pathogenesis of various disorders, such as obesity, diabetes, inflammatory conditions, and certain cancers.

2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) is an exogenous ligand that binds to the aryl hydrocarbon receptor (AHR), eliciting diverse toxic effects by disrupting the circadian clock mechanism. To investigate whether TCDD-activated AHR disrupts the intrahepatic circadian clock by interfering with genome-wide CLOCK:BMAL1 binding, potentially leading to a reduction or loss of rhythmicity in clock-controlled genes, interpretable machine learning models were developed to predict BMAL1 binding to DNA in liver, kidney, and heart tissues using genetic and epigenetic features (binding sequence, DNA shape, and histone modifications). Thus, TCDD activated AHR has been proposed to bind to the E-box binding motifs to disrupt the regulation of circadian clock genes. The findings demonstrated that BMAL1 binding to DNA is tissue-specific, and the combination of sequence, DNA shape, and histone modification features yielded the highest binding prediction accuracy. Additionally, the flanking sequences upstream and downstream of the binding motifs played a crucial role in BMAL1 binding to DNA.

Furthermore, a spatiotemporal multicellular mathematical model of the mammalian circadian clock in the liver lobule was developed to investigate intercellular coupling for the synchronization

of circadian clock expression across the portal-to-central axis. The analysis revealed that, similar to the coupling of autonomous circadian oscillators in the suprachiasmatic nucleus (SCN), hepatic clock rhythms are likely synchronized by an unknown coupling factor. Sensitivity analysis, bifurcation analysis, and parameter estimation from the model provided insights into the physiology of the hepatic clock and potential mechanisms of alteration.

Lastly, to understand the interplay between the spatial and temporal axes of gene expression in the liver, particularly in drug metabolism pathways, the existence of a third axis, chemical perturbation, and its implications for hepatic function were uncovered. We developed a non-linear mixed effect models to investigate the effect of acute TCDD perturbation on the spatial and temporal axes of gene expression in the liver lobule. The analysis revealed a distortion of the spatial axis of gene expression but a low significant effect on the temporal axis. These findings provide a comprehensive examination of circadian rhythms and their disruption by TCDD in the liver, encompassing molecular mechanisms, predictive modeling, and spatiotemporal dynamics. The study offers valuable insights into the intricate regulatory mechanisms governing circadian rhythms, the significance of zonation in hepatic functions, and the interplay between spatial and temporal gene expression. The findings have the potential to contribute significantly to our understanding of circadian resilience and the mitigation of pathological conditions, particularly in the context of drug metabolism pathways and hepatic function.

Copyright by
DANIEL KWABENA MARRI
2024

**I dedicate this thesis with profound gratitude and love to my family, all those who love me
and whom I love dearly.**

ACKNOWLEDGEMENTS

I would like to extend my sincere appreciation to my thesis supervisor and mentor, Dr. Sudin Bhattacharya, whose unwavering support, encouragement, and belief in me have been instrumental in shaping this endeavor. Your wealth of knowledge, patience, understanding, and provision of this wonderful opportunity have been instrumental in guiding me through this academic path. I also extend my heartfelt thanks to the esteemed members of my dissertation committee: Dr. Ripa Arora, for your guidance and the invaluable experience of introducing me to wet laboratory work; Dr. Dana Spence, for your steadfast support, teachings, and insightful advice; Dr. Hanne Hoffmann, for your constant availability, particularly in the early stages of my thesis, to discuss the biological relevance of my work – your research papers and conference recommendations were instrumental in propelling me this far; and Dr. Eva Farre Prokosch, for the honor of having you on my committee and for your insightful feedback that enriched the quality of my work. Additionally, I would like to extend my gratitude to Dr. Timothy Zacharewski, Dr. Rance Nault, and Dr. Cholino Giovanni for generously opening their doors and sharing their datasets with me; Dr. Pawel Danielewicz, for your support and introduction to Michigan State University; and Dr. David Arnosti, for being an exceptional mentor and teaching me the fundamentals of molecular biology, as well as for all the remarkable work you do for AIMS. Next, I would like to express my heartfelt appreciation to my family and friends who have been by my side every step of the way. Special thanks to Chido Matsika for your unwavering presence and support during the most challenging times, making them more manageable. I am grateful to Dr. David Filipovic, Dr. Omar Kana, and all my lab members for being a part of this significant journey and contributing to the completion of this thesis.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: PREDICTING MAMMALIAN TISSUE SPECIFIC DNA-BINDING BY CLOCK-BMAL1.....	8
INTRODUCTION.....	8
METHODS	13
RESULTS.....	17
CHAPTER 3: MATHEMATICAL MODELING OF THE SPATIAL AND TEMPORAL DYNAMICS OF CIRCADIAN CLOCK IN THE LIVER.....	37
INTRODUCTION.....	37
METHODS	41
RESULTS	51
CHAPTER 4: SPATIAL-TEMPORAL PERTURBATION OF THE LIVER LOBULE BY 2,3,7,8-TETRACHLORODIBENZO-<i>P</i>-DIOXIN	73
INTRODUCTION.....	73
METHODS	77
RESULTS.....	89
DISCUSSION AND CONCLUSION	127
BIBLIOGRAPHY.....	137

CHAPTER 1: INTRODUCTION

Circadian clocks are endogenous molecular networks present in the cells of wide spectrum of organisms from prokaryotes to eukaryotes that regulate 24 -h physiological and behavior rhythms¹⁻³. The circadian clock mechanism synchronizes physiological activities with the external cue cycles for organismal survival. In 1729, Jean-Jacques d'Ortous de Mairan discovered that *Mimosa pudica* leaves exhibit daily folding and unfolding cycles in constant darkness⁴. Animal circadian rhythms were later affirmed by experiments in 1972, which demonstrated persistent rhythmic activity in rats under constant darkness and temperature⁵ and in *Drosophila melanogaster* and other insects⁶. The circadian clock rhythms are regulated by a molecular clock mechanism which exists in all cells in the mammalian body. The hypothalamic suprachiasmatic nucleus (SCN) contains the master pacemaker clock that synchronizes subsidiary oscillators in peripheral tissues like the liver, heart, and kidney. The SCN receives input from external cues and, in turn, relays temporal signals to synchronize peripheral clocks^{2,7}. The molecular basis for the mammalian circadian clock mechanism involves a coordinated transcriptional-translational feedback loop that operates on an approximately 24 hr period. The feedback loop is driven by both positive and negative regulatory interactions that allow for sustained oscillatory activity⁸⁻¹⁰. The positive arm of the loop centers on the activation of E-box mediated transcription by the circadian locomotor output cycles kaput (CLOCK) or Neuronal PAS domain protein 2 (NPAS2) with brain and muscle ARNT-like 1 (BMAL1) forming the heterodimeric complex CLOCK-BMAL1 or NPAS2-BMAL1. This heterodimer protein drives the expression of various clock-controlled genes (CCG's) including the *Period* (*Per 1*, *Per 2* and *Per 3*) and the *Cryptochrome* (*Cry 1* and *Cry 2*) which comprise the negative arm by repressing their own transcription by interacting with CLOCK-BMAL1 or NPAS2-BMAL1 complex, thus constituting an auto-regulatory feedback process. Additionally, a

secondary stabilizing loop involves the activation of Bmal1 by retinoic acid receptor-related orphan nuclear receptors (ROR α , ROR β , and ROR γ) and its repression by REV-ERB α and REV-ERB β . The precise balance and delays involved in the kinetics of these multiple interconnected feedback processes allows for sustained 24-hour periodicity.

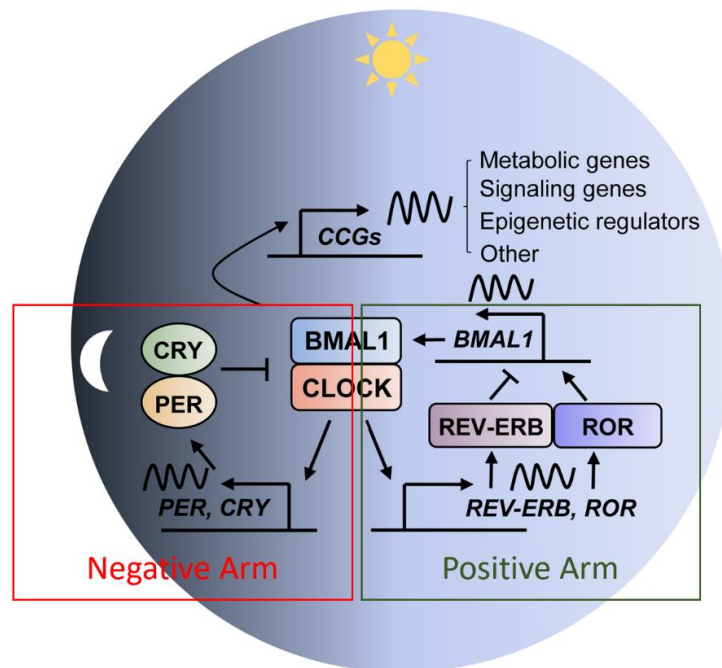


Figure 1: The key molecular component of the mammalian clock mechanism with the transcriptional – translational feedback loops driving by the positive and negative loop. Adapted from Lee, Y¹¹.

Several physiological and biological functions and processes are regulated by circadian clock-controlled genes^{12,13}, display rhythmicity over time. These include core body temperature, which fluctuates on a circadian cycle; sleep-wake cycles, which follow a circadian rhythm dictated by the suprachiasmatic nucleus; cardiovascular variables like heart rate and blood pressure, which demonstrate both circadian and ultradian rhythms; feeding behavior and digestive processes, which are influenced by circadian, hunger-satiety, and digestive cycles; secretion of hormones like

melatonin, cortisol, growth hormone, and prolactin, which often follow circadian or ultradian patterns determined by the hypothalamic-pituitary axis and other oscillators; liver metabolic activity and gene expression, which varies cyclically over 24-hour periods and in response to feeding times and fasting; renal blood flow, glomerular filtration rate, and other measures of kidney function, which display circadian variations; as well as several other measurable bodily processes

14–18.

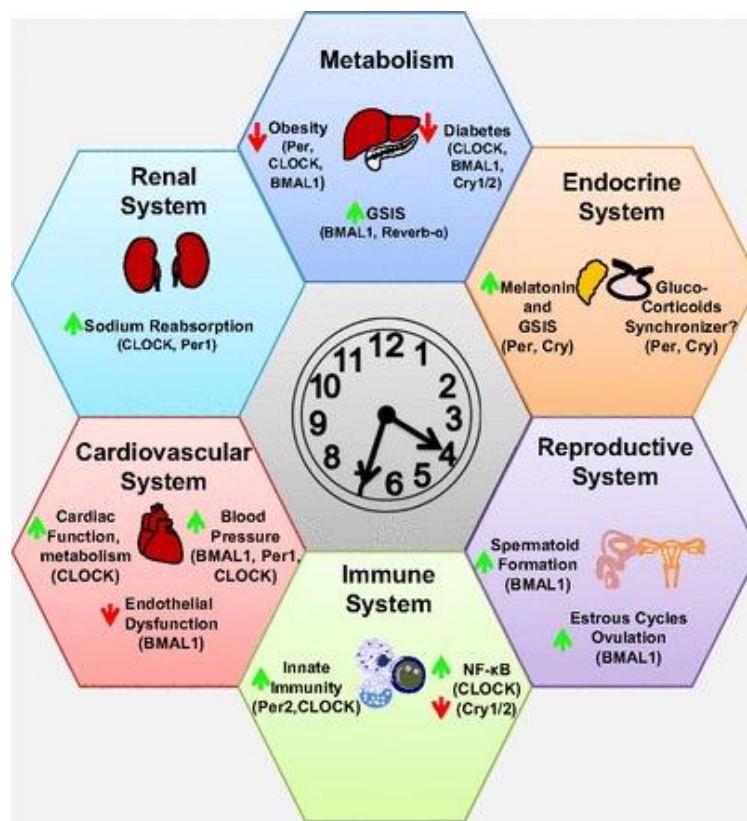


Figure 2: Physiological and biological functions regulated by circadian along with the implicated genes and proteins in parentheses. Green arrows represent induction and red arrows repression by circadian clock gene or protein. Adapted from Jacob et al¹⁹.

Studies from animal models and human subjects have shown that perturbation or desynchronization of circadian rhythms can contribute to, the progression of various pathologies^{20–22}. These include metabolic diseases like obesity and diabetes, inflammatory disorders, and certain cancers^{23–25}. Additionally, times of day can affect the severity of symptoms or acute exacerbations in some conditions with a circadian component. For example, asthma attacks and rheumatoid arthritis flares display circadian patterns, as does the incidence of adverse cardiovascular events like myocardial infarction and stroke^{18,24}. The cyclic nature of these conditions points to circadian misalignment and loss of coordination between external timing cues and internal oscillators as an integral mechanism in disease pathology. Strengthening circadian resilience through timed behavioral, or pharmacologic, or genetic interventions represents a therapeutic opportunity to mitigate disease severity in select disorders.

The chemical compound, 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) is a persistent organic pollutant that elicits diverse toxic effects in mammals. TCDD exposure has been associated with a myriad of adverse health outcomes, including developmental abnormalities, carcinogenesis, and hepatotoxicity. It exerts these effects primarily through activation of the cytosolic protein aryl hydrocarbon receptor (AHR), acting as a transcription factor^{26,27}. TCDD exposure perturbs circadian rhythms in peripheral tissues like the liver by disrupting molecular circadian clock oscillations. Specifically, TCDD activated AHR affects circadian rhythm by reducing the amplitude, shifting the phase, and lengthening the period of circadian oscillations in clock-controlled genes expression²⁸.

The liver exhibits strong rhythmicity, regulating various metabolic functions like, detoxification, glycogen storage, lipid metabolism, and bile production in accordance with the body's internal clock²⁹. The mammalian liver is made up of structural units called lobules. The lobule is composed

of hepatocytes, the liver's parenchymal cells, arranged in a radial pattern around a central vein. The hepatocytes are arranged in a roughly hexagonal architecture with a pericentral to periportal axis extending from the central to the portal veins. Hepatocytes are not homogeneous in their metabolic activities; instead, they display distinct biological functions depending on their location within the lobule³⁰⁻³². Thus, metabolic functions within the liver lobule exhibit a distinct spatial organization, a phenomenon known as zonation. The periportal hepatocytes, being exposed to a higher oxygen concentration, are primarily involved in processes such as gluconeogenesis (the synthesis of glucose from non-carbohydrate precursors), urea synthesis (detoxification of ammonia), and bile acid synthesis. Conversely, the pericentral hepatocytes, which receive oxygen-depleted blood, are more active in glycolysis (the breakdown of glucose for energy production), lipogenesis (the synthesis of fatty acids), and xenobiotic metabolism (the detoxification of drugs and other foreign compounds)^{30,32-36}.

The hepatic expression of AHR is zoned across the lobular axis^{29,33}. However, no zonation is observed in the basal expression of circadian clock genes. While a large number of mathematical models of the natural circadian clock oscillation have been developed^{9,37,38}, none of these specifically address spatial disruption of the circadian clock in peripheral tissues like the liver. I aimed to address the question of whether TCDD-activated AHR disrupts the intra-hepatic circadian clock in a zoned manner by interfering with genome-wide CLOCK:BMAL1 binding, which could lead to a reduction or total loss of rhythmicity of clock-controlled genes. To do this, I first developed an interpretable machine learning model of genome-wide DNA binding by BMAL1 in the liver, heart and kidney using published Chromatin immunoprecipitation-sequencing (ChIP-seq) data. My results confirmed an overlap between the binding motif of BMAL1 (E-box – 5'-CANNTG-3') and AHR (5'-GCGTG-3') in the promoter regions and bodies of core circadian clock

genes. This is consistent with the proposed mechanism of TCDD activated AHR binding to the E-box motif to disrupt the regulation of circadian clock genes. Also, my model showed that, using both genomic and epigenomic features like the core motif plus flanking sequences, the shape of the DNA and histone modification, we could achieve significant prediction accuracy of BMAL1 binding. Interpreting my model, I showed specific features with the highest contribution to tissue-specific binding of BMAL to DNA. My cross tissue predictive model demonstrated that while BMAL1 exhibits high specificity in binding certain DNA conformations and chromatin contexts, these specificities exhibit variation across different tissues. Our discoveries expand upon the notion that DNA shape and chromatin context can modulate the binding specificities of transcription factors (TFs). Specifically, our results demonstrate that in addition to conferring differential binding specificities to distinct TFs within the same tissue, DNA shape and chromatin environment can also confer distinct binding specificities to a given TF in different tissue contexts.

Secondly, I developed spatiotemporal multicellular mathematical models of the mammalian liver lobule circadian clock network. I developed two sets of mathematical models to investigate the dynamics of circadian clock genes expression in hepatocyte across the portal-to-central axis of the liver lobule: one model with communication and one without communication among cellular oscillators (hepatocytes). Simulations of the model revealed the dependencies of the model observables (expressed mRNA levels of the various genes) on their corresponding transcription and degradation rates. Estimating the model parameters and fitting the coupled model to experimental data yielded a high correlation with $R^2 > 0.9$, elucidating the way alterations in model parameters modulate the reinforcing and attenuating feedback loops that govern the circadian rhythm. Collectively, my modeling framework establishes a foundation for probing the regulatory

mechanisms underlying circadian rhythmicity and its associated aberrations in the context of hepatic physiology and pathological states.

A recent investigation by Droin et al²⁹ revealed the intricate interplay between spatial and temporal axes of gene expression, leading to rhythmic patterns of expression within established zoned pathways. Notably, these pathways encompass a significant number of gene sets implicated in drug metabolism processes. The observed overlap between the temporal and spatial axes with drug metabolism pathways suggests the existence of a third axis, chemical perturbation, that warrants consideration when characterizing hepatic function. This led to the final project in my dissertation where I developed a nonlinear mixed effect model to investigate the effects of acute TCDD-activated AHR on the temporal (rhythmicity) and spatial (zoned) expression patterns of genes along the portal-central axis of the liver lobule.

CHAPTER 2: PREDICTING MAMMALIAN TISSUE SPECIFIC DNA-BINDING BY CLOCK-BMAL1

INTRODUCTION

All living organisms possess a robust circadian timekeeping mechanism enabling anticipation and adaptation to recurring environmental changes. In mammals, this system consists of a hierarchical network of oscillators. The central clock, situated in the suprachiasmatic nucleus (SCN) of the hypothalamus, synchronizes peripheral clocks across various tissues³. The intracellular gene regulatory network of the circadian clocks, both central and peripheral, involves a relatively small set of key transcription factors (TFs) that are interconnected through multiple negative and positive feedback loops. These feedback loops play a crucial role in regulating the expression of clock genes and maintaining the oscillatory patterns of the circadian rhythm. In this regulatory network, specific TFs bind to the promoter regions of target genes, either activating or repressing their transcription. The expression of these TFs is, in turn, regulated by the products of the genes they control, creating intricate feedback loops³⁹. The core activators of the circadian network are, the Clock Locomotor Output Cycles Kaput (CLOCK) and brain and muscle ARNT Like 1 (BMAL1) transcription factors, both of which belong to the basic helix-loop-helix (bHLH) family. These two proteins form a heterodimer complex (referred to hereon as CLOCK-BMAL1). In the absence of CLOCK, another member of the bHLH-PAS transcription factor family, the Neuronal PAS domain protein 2 (NPAS2), can compensate by forming a heterodimer with BMAL1⁴⁰. In the classical model of clock gene regulation, the CLOCK-BMAL1 or NPAS2-BMAL1 dimer acts as a transcriptional activator, initiating the expression of various clock-controlled genes by binding to a specific hexanucleotide sequence known as the E-box motif (canonical sequence 5'-CANNTG-3', where N represents any nucleotide) within the promoter or enhancer regions of clock-controlled genes. This binding event regulates the transcription of these genes, which are crucial for the proper

functioning of the circadian clock machinery^{40,41}. In vivo studies have shown that, BMAL1 binds also to non-canonical E-box sequences such as 5'-CACGTT-3' in the promoter region of the murine *Per2* gene⁴². However, comprehensive experimental evidence supporting genome-wide binding of BMAL1 to such sequences remains elusive. Consequently, in this study, I have focused solely on the classical E-box motif with the canonical sequence 5'-CANNTG-3' (where N represents any nucleotide). Disruptions in the expression or binding functionality of the core clock transcription factors (TFs) disturb natural circadian oscillations and can lead to various pathologies, including insomnia, cancer, cardiovascular disease, and metabolic disorders^{43,44}. Here, I aim to enhance our understanding of gene regulation by the CLOCK-BMAL1 or NPAS2-BMAL1 complex and its perturbations by employing interpretable predictive models of DNA binding by the master regulatory factor BMAL1. By developing these models, I seek to gain insights into the mechanisms underlying the binding of BMAL1 to target DNA sequences and the subsequent regulation of clock-controlled genes.

The genome-wide identification of transcription factor binding sites (TFBS) poses a significant challenge. Typically, only a small fraction of the classically defined sequence motifs for a particular transcription factor (TF) are bound⁴⁵. For example, the canonical E-box binding motif occurs more than 7 million times across the mouse genome, but less than 0.7% of these motifs are bound by the CLOCK-BMAL1 or NPAS2-BMAL1 complexes in mouse peripheral tissues⁴⁶. The binding of a particular TF to its cognate DNA motif depends on several molecular features, including the DNA sequence of the core motif, sequences flanking the core motif, chromatin accessibility, local shape of the DNA, presence of co-factors, histone modifications, DNA methylation, and other biophysical parameters⁴⁷⁻⁵⁰. These features and their relative contributions to binding can vary greatly across different cell and tissue types^{51,52}. Chromatin

immunoprecipitation followed by sequencing (ChIP-seq) is the current gold standard for assaying genome-wide TF binding locations⁵³. However, assaying the binding of a given TF under various conditions and in different tissues is prohibitively expensive. Consequently, several predictive computational models of genome-wide TF-DNA binding have been developed. From these models, DNA sequence and chromatin accessibility emerge as the most important determinants of TF binding. Despite the vast number of potential binding sites in the genome, TFs exhibit highly specific binding patterns, with only a small fraction of the canonical motifs being occupied. This specificity is attributed to the interplay of various factors, including the DNA sequence context, chromatin accessibility, and other biophysical parameters. These features can vary significantly across different cell and tissue types, contributing to the complexity of TF binding patterns⁵⁴⁻⁵⁶. Chromatin accessibility assays, such as deoxyribonuclease hypersensitive sites sequencing (DNase-seq) and assay for transposase-accessible chromatin sequencing (ATAC-seq), have been employed to enhance the prediction of transcription factor binding sites (TFBS). These assays provide valuable information about the regions of the genome that are accessible to regulatory proteins, a crucial determinant of TF binding⁵⁷. Recently, advancements in machine learning, particularly deep learning techniques, have led to improved model predictions for TF binding. Deep learning models have demonstrated the ability to capture complex patterns and interactions within the data, leading to more accurate predictions of TF-DNA binding events. While deep learning models excel in predictive performance, they often tend to be "black boxes," where it is challenging to understand the reasoning and mechanism behind their predictions. This lack of interpretability hinders the ability to elucidate the molecular features and interactions that contribute to the tissue-specific binding patterns of transcription factors⁵⁸⁻⁶⁰.

In this study, I present interpretable machine learning models capable of predicting which canonical E-box motifs occurring in accessible chromatin regions of the mouse liver, heart, and kidney are likely to be bound by BMAL1. Our predictive models are based on the XGBoost⁶¹ machine learning algorithm, with logistic regression used as a baseline algorithm to evaluate model performance. Published data from a BMAL1 ChIP-seq study⁴⁶ was used to train and evaluate the models. When considering which features to include in our predictive models, I noted that DNA shape^{62,63} and histone modifications⁶⁴ have been shown to be efficient predictors of transcription factor (TF) binding in addition to DNA sequence. Specifically, it has been proposed that TFs prefer specific 3D DNA conformations and not just specific sequences⁶⁵. For example, the incorporation of DNA shape features led to improved model performance when predicting *in vivo* binding of TFs from the basic helix–loop–helix (bHLH) family. Particularly, five distinct shape features - Electrostatic Potential (EP), Minor Groove Width (MGW), Propeller Twist (ProT), Roll, and Helix Twist (HelT) have been shown to be useful for TF-DNA binding prediction. These shape features capture the local three-dimensional structure of the DNA, which can influence the binding affinity and specificity of TFs⁶⁶. In addition to DNA shape features, histone modifications have also been implicated in regulating TF-DNA binding. Specific histone modifications can alter the chromatin structure and accessibility, thereby affecting the ability of TFs to recognize and bind to their target sequences. By incorporating information about histone modifications, predictive models can potentially capture additional regulatory mechanisms governing TF-DNA binding⁶⁷.

Our study revealed that while most of the flanking DNA sequence features showed low importance in predicting the binding of BMAL1 to canonical E-box motifs. However, the second flanking nucleotide upstream of the E-box motif in the liver, specific histone modifications and DNA shape features emerged as significant predictors of BMAL1-DNA binding across all tissues examined.

Specifically, the histone modifications H3K27ac, H3K4me1, H3K4me3, and H3K36me3, together with the DNA shape features electrostatic potential (EP), roll, and minor groove width (MGW), were identified as crucial determinants of BMAL1-DNA binding in the liver, heart, and kidney tissues. The incorporation of these features into our machine learning models resulted in high predictive performance, highlighting their importance in governing the binding specificity of BMAL1. However, our cross-tissue predictive model revealed that although BMAL1 exhibits a preference for specific DNA conformations and chromatin contexts, these specificities vary across different tissues. This finding suggests that while certain genomic and epigenomic features are generally important for BMAL1 binding, the relative contributions of these features and the specific combinations that facilitate binding may differ among tissues. The tissue-specific variations in BMAL1 binding preferences could be attributed to the unique chromatin landscapes, regulatory networks, and cellular environments present in different tissues. These differences may influence the interplay between DNA sequence, DNA shape, histone modifications, and other regulatory factors, resulting in distinct binding patterns of BMAL1 across tissues.

Our interpretable machine learning models not only achieved high predictive accuracy but also provided insights into the key genomic and epigenomic features that govern BMAL1-DNA binding. By identifying the most influential features and their relative importance, our study contributes to a better understanding of the regulatory mechanisms underlying the tissue-specific binding patterns of BMAL1, a critical transcription factor in the circadian clock machinery.

METHODS

ChIP-seq dataset preprocessing.

Uniformly processed BMAL1 ChIP-seq peaks from the C57BL/6J mouse liver, kidney, and heart were retrieved from the Gene Expression Omnibus under accession code GSE1106049⁴⁶. BMAL1 ChIP-seq experiments were conducted at Zeitgeber time 6 (ZT6). Accessible chromatin regions, represented by DNase I-hypersensitive (DHS) sites for all three tissues (DNase-seq), were obtained from the Encyclopedia of DNA Elements, ENCODE (Supplementary Materials). DNase-seq experiments were performed on unsynchronized tissues. The Genome Reference Consortium Mouse Build 38 (GRCm 38) served as the reference genome. DHS sequences were processed in Python with BEDTools⁶⁸ to extract all E-Box sequences (5'-CANNTG-3') within accessible chromatin. E-box motifs in accessible chromatin regions not overlapping their respective tissue ChIP-seq bed files were considered instances of unbound motifs (the negative dataset for the model). All accessible chromatin singleton E-boxes (instances of only one E-box motif under a BMAL1 peak) and E-boxes closest to the summit of the BMAL1 peaks for peaks with multiple E-boxes were labeled as bound (the positive dataset). Other E-boxes under BMAL1 peaks were deemed ambiguous and excluded from further analysis. Specifically, 1175 E-boxes, 1082 E-boxes, and 663 E-boxes from the bound Bmal1 liver, kidney, and heart, respectively, were found to be ambiguous due to multiple E-boxes. Each E-box motif sequence was extended to include 4-basepair (bp) flanking sequences upstream and downstream of the E-box. As the E-box motif sequence is a palindrome, the reverse complement was disregarded. Each E-box, thus represented by a 14-nucleotide sequence (6-bp core plus 4-bp sequence on either end), was one-hot encoded. The binary (bound and unbound) E-box data resulted in highly imbalanced datasets, with significantly more unbound than bound E-boxes in mouse accessible chromatin. Specifically, the

number of bound E-box motifs in the liver, kidney, and heart were 3725, 3237, and 1313, respectively. Conversely, the number of unbound E-box motifs in the liver, kidney, and heart were 189581, 262053, and 291840, respectively. Notably, the negative samples outnumbered the positives by factors of 51 in the liver, 223 in the heart, and 82 in the kidney. The reported occupancies and the ratio of bound to unbound E-box motifs in the liver, kidney, and heart tissues are consistent with previous studies, indicating a low percentage of canonical E-box motifs being bound by BMAL1 in accessible chromatin regions.

DNA shape preprocessing.

The DNA sugar-phosphate backbone possesses degrees of freedom, allowing neighboring base pairs and bases within a pair to vary their positions relative to each other through rotation or translation, resulting in changes to the overall shape of the DNA molecule. To estimate DNA shape features, the R/Bioconductor package DNASHapeR^{69,70} was employed. The DNASHapeR algorithm predicts DNA shape features based on a given DNA sequence and encodes them into feature vectors. These feature vectors for each shape category were normalized to values between 0 and 1 using Min-Max normalization and grouped into sets of 10 values for Minor Groove Width (MGW), Propeller Twist (ProT), and Electrophoretic Mobility (EP), and sets of 11 values for Helix Twist (HelT) and Roll, to serve as inputs for predictive models. The number of bins for each shape feature corresponds to the length of the sliding window used to generate the features – 5 base pairs (bp) for MGW, ProT, EP, and 6 bp for HelT and Roll.

Histone modification preprocessing.

I obtained ChIP-seq data for five histone modifications, namely H3K27ac, H3K4me1, H3K4me3, H3K27me3, and H3K36me3, in mouse liver, kidney, and heart tissues from the ENCODE database (Supplementary Materials). The ChIP-seq experiments for histone modifications were performed

on unsynchronized tissue samples. These specific histone modifications were chosen based on data availability across all tissues and their established roles in transcription factor binding. The corresponding bed files were utilized to generate signal profiles and heatmaps using the deepTools software⁷¹. From the generated profiles and heatmaps, we observed that the histone modification ChIP-seq signals extended meaningfully up to a 1.5-kb region (+/- 750 bp) centered on the E-box core motif. Focusing on this 1.5-kb region centered at the E-box core motif, we extracted the histone modification features for the binary dataset for each tissue using the bwtool software⁷². Subsequently, these features were divided into 10 bins, with an equal number of nucleotides in each bin.

Machine learning models.

XGBoost

Extreme Gradient Boosting (XGBoost) is an ensemble learning method based on boosting decision trees for both classification and regression tasks⁶¹. I used up to 20 features as inputs for each E-box motif - 10 sequence features (one for each nucleotide), 5 DNA-Shape features and 5 histone modification features. The first two and last two nucleotide of the E-box motifs were set because they were the same in all motifs. Using the Scikit-learn library in Python, I performed hyperparameter tuning of the following parameters to reduce the degree of overfitting - the number of iterations in training (`n_estimators`), the sum of sample weight of the smallest leaf nodes to prevent overfitting (`min_child_weight`), the maximum depth of the tree in building a model while training (`max_depth`), the sampling rate of the training set in each iteration (`subsample`), the learning rate (`learning_rate`), and the feature sampling rate when constructing each tree (`colsample_bytree`). Hyperparameter tuning of the XGBoost model involved exploring a grid search of the hyperparameter space with specific values: `n_estimators` = {30, 40, 50, 60, 70, 80,

90, 100}, min_child_weight = {1, 2, 3, 4, 5, 6}, subsample={0.5, 0.6, 0.7, 0.8, 0.9, 1 }, max_depth = {1, 2, 3, 4, 5}, learning_rate = {0.1, 0.2, 0.3, 0.4, 0.5}, and colsample_bytree = {0.6, 0.7, 0.8, 0.9, 1 }, resulting in a potential combination of 36,000 hyperparameters. Additionally, I evaluated the model's performance using 5-fold cross-validation on predicting the binding status of E-box motifs in accessible chromatin regions.

Logistic Regression

Logistic regression is a parametric classification model that estimates the probability of the output variable belonging to a particular class⁷³. It serves as a baseline for most machine learning-based classification models. In this study, I tuned the following hyperparameters of the logistic regression model to reduce overfitting on our testing dataset: the regularization solver for the training dataset (solver) and the maximum number of iterations the solver algorithm is allowed to run before convergence (max_iter). The regularization solver determines the algorithm used for optimization and regularization. Common solvers include 'lbfgs' (limited-memory Broyden-Fletcher-Goldfarb-Shanno), 'newton-cg' (Newton-Conjugate Gradient), and 'liblinear' (coordinate descent algorithms). The max_iter parameter specifies the maximum number of iterations the solver is permitted to run before terminating the optimization process. By tuning these hyperparameters, I aimed to find the optimal combination of solver and maximum iterations that would minimize overfitting on the testing dataset, thereby enhancing the generalization performance of the logistic regression model.

RESULTS

BMAL1 binds most frequently to the CACGTG E-box motif in all tissues.

To understand the molecular factors governing the binding of BMAL1 to DNA, I first delved into the role of the canonical 5'-CANNTG-3' E-box motif in the binding of BMAL1 to DNA in the various tissues. I conducted a comprehensive scan of the mouse mm10 reference genome to identify occurrences of the E-box motif; 5'- CANNTG -3'. I considered all possible nucleotide permutations of the central two nucleotides (NN) and included the reverse complement of the canonical E-box sequence. However, a particular E-box and its reverse complement were treated separately. Utilizing DNase-seq datasets retrieved from the ENCODE database⁷⁴ for C57BL/6J mouse tissues (liver, heart, and kidney), I identified subsets of E-boxes that intersected or overlapped with DNase-seq hypersensitive sites (DHS), indicative of accessible chromatin. The tissue-specific list of E-boxes from accessible chromatin were then compared with their respective tissue matched BMAL1 ChIP-seq⁴⁶ peaks to extract all BMAL1 bound and unbound E-Boxes in accessible chromatin. Furthermore, I identified instances where BMAL1-bound E-boxes were situated outside of accessible chromatin, accounting for 0.8% of all peaks. To prevent potential confounding effects between the two classes of bound E-boxes, I excluded these instances from our model training and evaluation processes. Comparison of BMAL1-bound E-box occurrences within accessible chromatin across liver, heart, and kidney tissues revealed marked tissue specificity. Specifically, only 398 E-boxes were found to be bound in common across all three tissues (Fig 3A-C). E-boxes bound across all three tissues were frequently located within the promoters of core circadian clock genes, as indicated by our findings (data not shown).

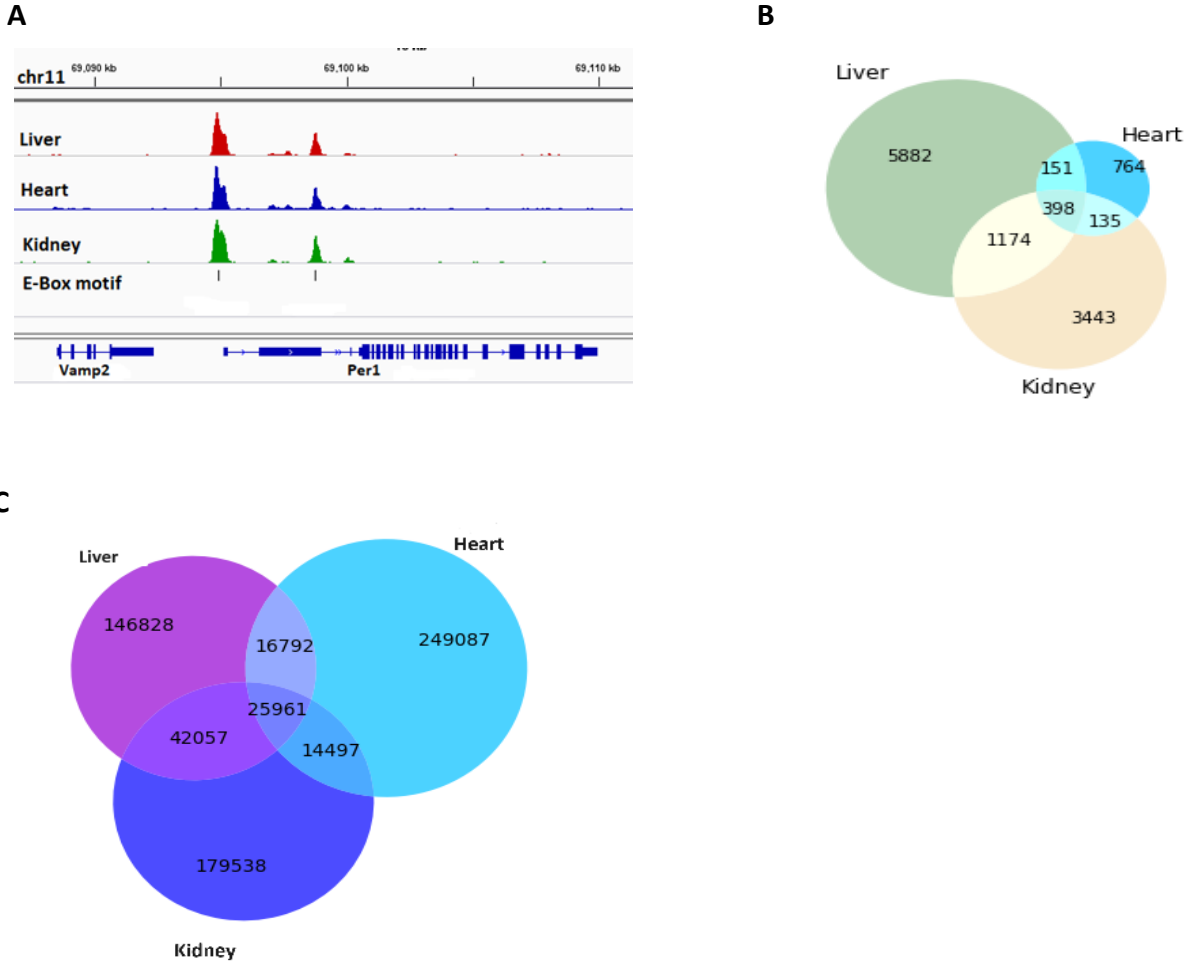


Figure 3: **(A)** BMAL1 ChIP-seq peaks in the liver (red), kidney (blue), and heart (green), and E-box binding motifs (black vertical bars) under the peaks at the *Per1* locus. **(B)** Venn diagram representing the overlap of bound E-boxes motifs in open chromatin across liver, kidney, and heart. **(C)** Venn diagram representing the overlap of unbound E-boxes motifs in open chromatin across liver, kidney, and heart. Adapted from Marri et al⁷⁵.

Subsequently, I quantified all occurrences of the canonical E-box motif (5'-CANNTG-3') throughout the mouse genome, wherein N denotes any nucleotide type. The canonical E-box encompasses 16 unique E-box types, corresponding to each permutation of the NN dinucleotide situated at the center of the motif. I computed the fraction of each individual E-box type compared to the total number of E-boxes (Fig 4A). Notably, the E-box types 5'-CACATG-3' and 5'-

CATGTG-3' collectively constituted the highest fraction of E-boxes within the mouse genome, accounting for 17.3% of all instances. These two motifs, being reverse complements of each other, demonstrated approximately equal frequencies, like all other non-palindromic E-boxes that exhibit reverse complementarity. Interestingly, the palindromic BMAL1-preferred E-Box motif, 5'-CACGTG-3', occurs the fewest number of times, representing only 1.83% of all instances within the mouse genome. Subsequently, I employed the same methodology to analyze E-boxes within accessible chromatin regions of mouse liver, kidney, and heart. The palindromic motif CAGCTG was the most common E-box type across accessible chromatin regions in all three tissues, while the BMAL1-preferred E-Box CACGTG was among the three least common motifs, which were all palindromes (Fig 4B). Using the overlap between the BMAL1 ChIP-seq and DNase-seq peaks, I computed the percentage of BMAL1 bound E-boxes in the mouse liver, kidney and heart relative to the total number of E-boxes of the same type within accessible chromatin of their respective tissues. Interestingly, the BMAL1-preferred E-box 5'-CACGTG-3' emerged as the most frequently bound E-box type across all three tissues. In addition, about 18% of CACGTG E-boxes accessible in the liver were bound, and for the kidney and heart these fractions were 15%, and 4%, respectively. Furthermore, less than 20% of all individual E-boxes identified within accessible chromatin in any given tissue were also bound in that same tissue (Fig 4C). The kidney and heart had a higher overall number of E-boxes within these open chromatin regions compared to the liver. However, despite having a lower total count of E-boxes in accessible chromatin, the liver exhibited a higher proportion of BMAL1-bound E-boxes. This observation suggests a differential regulatory landscape for BMAL1 binding among these tissues, potentially reflecting distinct functional roles and circadian regulatory mechanisms.

I observed instances where there were none (zero), exactly one (singleton), or two or more (multi) E-box motifs under a single BMAL1 ChIP-seq peak across all tissues examined (Fig 4D). I then extracted all singleton E-boxes and the E-boxes closest to the summit of the BMAL1 peak within multi-E-box peaks, categorizing them as bound peaks (positive dataset). The E-Boxes present in accessible chromatin regions that were not bound by BMAL1 were labeled as unbound peaks (negative dataset). All other E-boxes were excluded from further analysis. The ratios of the positive (bound) to negative (unbound) peaks were 1:51, 1:82, and 1:223 in the liver, kidney, and heart, respectively. These ratios indicate a significant imbalance between the bound and unbound E-boxes, with a substantially higher number of unbound E-boxes in all three tissues.

Together, these results suggest that BMAL1 likely interacts with multiple different E-box types across the liver, kidney, and heart in a tissue-specific manner. Notably, the E-box motif CACGTG was found to be the most highly associated with BMAL1 binding among the different E-box types analyzed, despite its lower occurrence in the genome overall. The observation of varying ratios of bound to unbound E-boxes across tissues highlights the tissue-specific nature of BMAL1 binding and suggests that additional factors, such as chromatin accessibility, co-factor availability, and regulatory mechanisms, may contribute to the differential binding patterns observed. Furthermore, the identification of instances where BMAL1 ChIP-seq peaks do not overlap with any E-box motifs or contain multiple E-boxes under a single peak underscores the complexity of BMAL1 binding and the potential involvement of alternative binding mechanisms or indirect interactions with chromatin.

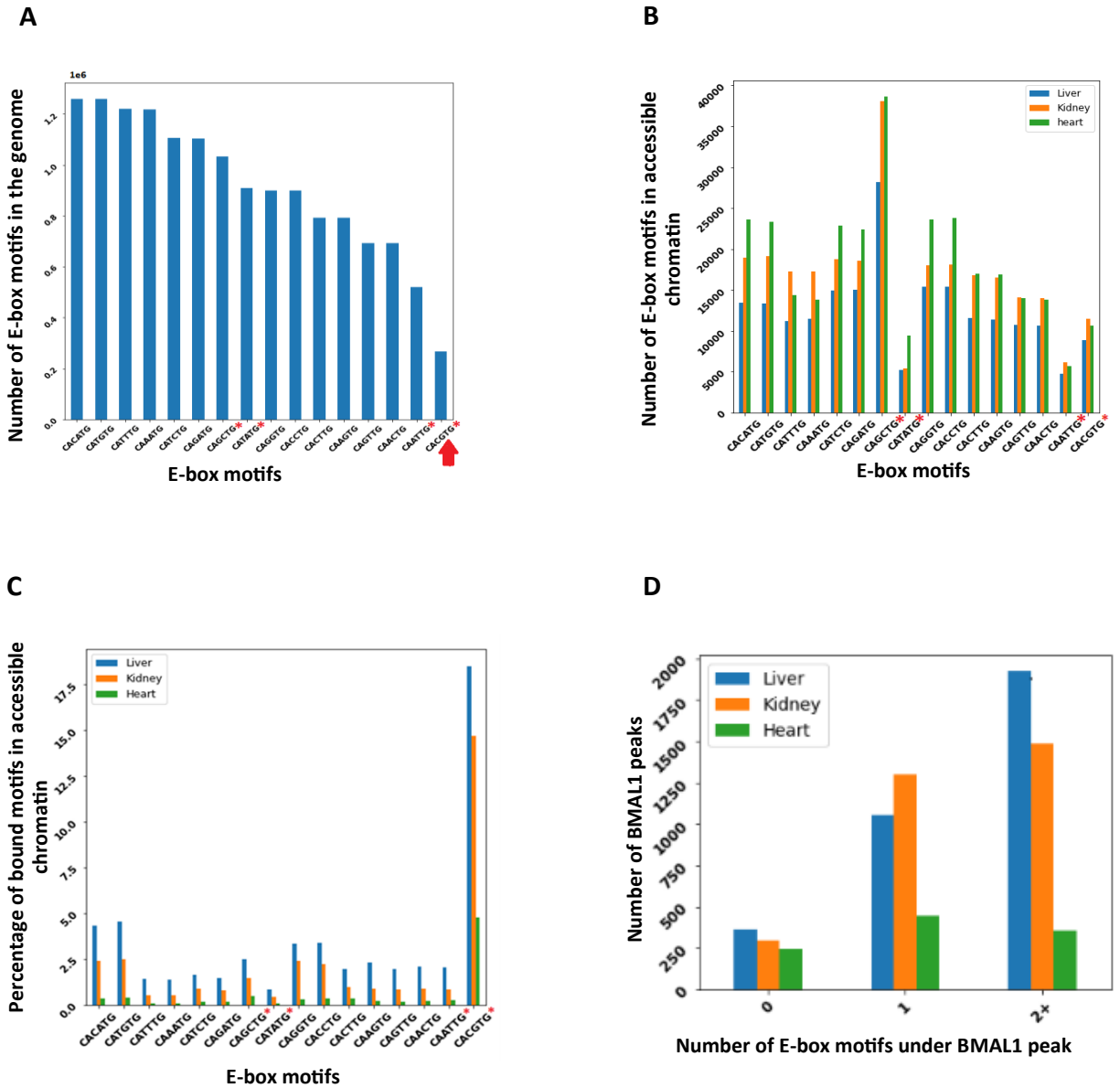


Figure 4: **(A)** E-box binding motif distribution across the entire mouse genome. The canonical E-Box motif CACGTG (marked with an arrow) is the least represented motif in the mouse genome. **(B)** Distribution of E-box binding motifs in open chromatin across the liver (blue), kidney (orange) and heart (green). **(C)** Percentage of BMAL1 bound E-box motifs in open chromatin across the liver (blue), kidney (orange) and heart (green). **(D)** Distribution of BMAL1 peaks with zero (0-E-Box), exactly one (singleton E-box) and multi (two or more E-box) E-box motifs in the liver (blue), kidney (orange) and heart (green). In (A), (B) and (C) the sequences along the x-axis are ordered by their frequency in the mouse genome shown in (A), which also happens to group complementary sequences adjacent to each other. The four palindromic sequences, which are their own complements, are marked with asterisks. Adapted from Marri et al⁷⁵.

Predicting genome wide BMAL1 binding within tissues.

Given that nucleotides flanking the E-box have been demonstrated to influence the binding specificity of E-box binding transcription factors (TFs)⁵⁵. Therefore, I extended and one-hot encoded the genomic sequence for all BMAL1-bound (positive) and unbound (negative) E-boxes by 4 base pairs (bps) upstream and downstream of the E-box. Additionally, I computed the following DNA shape features for the extended 14 bp sequence: Electrostatic Potential (EP), Minor Groove Width (MGW), Propeller Twist (ProT), Roll, and Helix Twist (HelT), utilizing the k-mer + k-shape (k=1) sequence feature model⁵⁰. Although the shape features are derived from the DNA sequence, they can potentially capture higher-order interdependencies between neighboring nucleotides, thereby providing additional information to the model input. DNA shape features can also elucidate the importance of flanking sequences in TF-DNA binding specificity⁵⁵. Visualization of the DNA shape features EP, ProT, and Roll revealed disparities in DNA shape between bound and unbound motifs across the liver, kidney, and heart, whereas the MGW feature exhibited differences between bound and unbound motifs solely in the kidney. The shape feature vector for each category was then normalized to values ranging from 0 to 1 using Min-Max normalization, grouped into intervals of 10 values for the DNA shape features EP, MGW, and ProT, and intervals of 11 values for HelT and Roll. These normalized DNA shape feature vectors were utilized as input features for the predictive models, as depicted in Fig 5.

Epigenetic modifications are known to influence transcription factor binding. Specifically, histone modifications intricately regulate transcription factor occupancy and subsequently modulate gene expression patterns^{62,76}. Histone modification ChIP-seq signal values, encompassing genomic regions spanning ± 750 base pairs (bps) around the E-box, were utilized to compute feature vectors for five histone modifications: H3K27ac, H3K4me1, H3K4me3, H3K27me3, and H3K36me3. The

selection of the ± 750 -bp region was guided by the need to consider local profiles of histone modifications, approximating the size of a typical promoter or enhancer. The histone feature vector was partitioned into 10 bins, with the signal strength averaged across 150 bps within each bin. This approach allowed for the comprehensive characterization of histone modification patterns surrounding the E-box, providing insights into the regulatory landscape governing transcription factor binding dynamics and gene expression regulation.

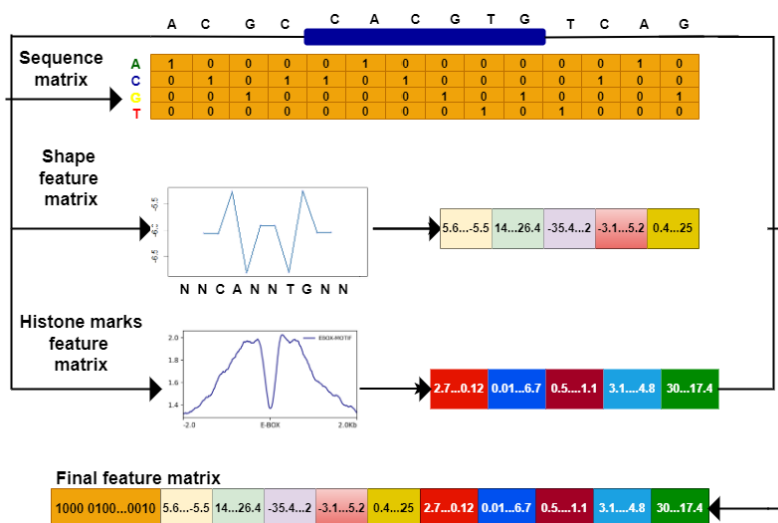


Figure 5: Design of the machine learning algorithm input features. The local chromatin features (E-box DNA sequence features) and flanking sequences were one-hot encoded. The DNA shape genomic feature matrix from the k-mer + k-shape (k=1) sequence feature model and epigenomic (histone modification) features averaged and binned were used as the final feature matrix for the model. Adapted from Marri et al⁷⁵.

I implemented three distinct models to investigate the predictive power of various feature combinations for determining the binding status of E-boxes in accessible chromatin regions. The models were constructed using subsets of the final encoded feature set, encompassing (i) DNA sequence information alone (DNA sequence-only model); (ii) a combination of DNA sequence and DNA shape features (sequence + shape model); and (iii) an integrated model incorporating

DNA sequence, DNA shape, and histone modification data (sequence + shape + HM model). To predict the binding status of E-boxes, I employed two machine learning algorithms: XGBoost (eXtreme Gradient Boosting) and logistic regression. XGBoost, a powerful tree-based ensemble learning algorithm, served as our principal predictive model, while logistic regression was utilized as a baseline for performance comparison. Aiming to optimize model performance, I employed a grid search strategy in conjunction with stratified 5-fold cross-validation to tune the hyperparameters of each model. This process involved partitioning the data into five stratified folds, where each fold was used as a validation set once, while the remaining four folds were used for training. Hyperparameter tuning was performed independently for each model based on the liver, heart, and kidney datasets, allowing us to derive the optimal hyperparameter configurations tailored to each tissue-specific dataset. Subsequently, the models with the optimized hyperparameters were trained through five-fold stratified cross-validation, and their predictive performance was evaluated for the liver, heart, and kidney datasets (Figure 6 and table 1). The average performance across the five folds was reported for each model and tissue type.

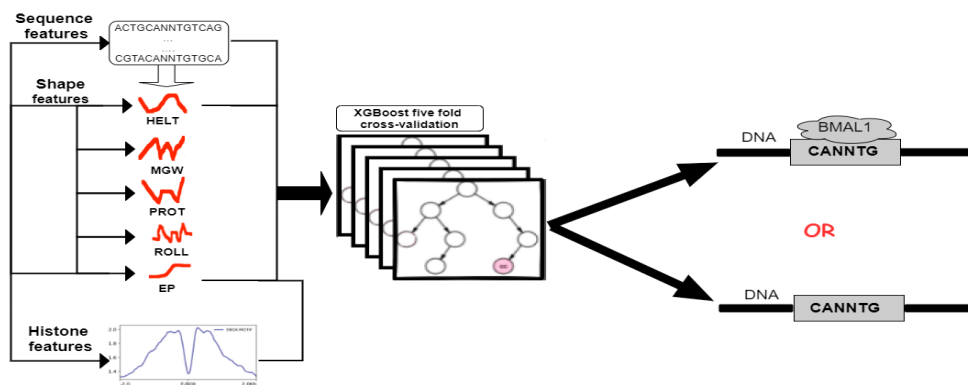


Figure 6: Schematic of machine learning- predictive model training. Based on 5-fold cross-validation, the XGBoost classifier predicted the binding status of E-box motifs in open chromatin, training on all accessible bound E-boxes and unbound E-boxes. Adapted from Marri et al⁷⁵.

DNA shape and histone modification features improve within-tissue model performance.

Model performance was assessed using two widely adopted metrics: the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). These metrics provide a comprehensive evaluation of the models' ability to distinguish between positive (bound) and negative (unbound) instances, accounting for both sensitivity and specificity.

Performance of sequence-only models: We conducted training and validation of our XGBoost classifier across liver, heart, and kidney datasets, utilizing 10 sequence features: the 2 central nucleotides of the E-Box, along with an additional 4 flanking nucleotides upstream and downstream of the E-box (NNNNCANNTGNNNN, excluding the conserved CA and TG subsequences) due to the repetition of CA and TG nucleotide in all the sequence features. To evaluate classifier performance, we computed the average AUROC and AUPRC scores for each tissue via stratified 5-fold cross-validation. Given the unbalanced distribution between the two classes - bound vs unbound E-boxes, AUPRC was deemed a more appropriate metric in our case. The mean AUROC scores across liver, kidney, and heart were 0.71, 0.78, and 0.80, respectively. Correspondingly, the mean AUPRC scores were 0.09, 0.10, and 0.06 for the liver, kidney, and heart, respectively (Fig 7A&B). The relatively high AUROC and AUPRC scores across all tissues suggest differences in the two central nucleotides and flanking sequence between BMAL1 bound and unbound E-boxes. However, the predictive power solely derived from DNA sequence seems insufficient for robust prediction, indicating the need for additional features or more sophisticated modeling approaches to enhance prediction accuracy.

Performance of sequence + shape models: The specific local configurations of DNA are determined by its three-dimensional structure, which in turn influences various biological

processes. Computational methods, including Monte Carlo simulations, have been utilized to derive features that quantify DNA shape from its local sequence^{66,69}. Five distinct DNA shape features – Electrostatic Potential (EP), Minor Groove Width (MGW), Propeller Twist (ProT), Roll, and Helix Twist (HelT) – have been identified as significant contributors to the DNA binding affinity of transcription factors belonging to the basic helix-loop-helix (bHLH) family⁵⁰. The 14 bp sequences (NNNNCANNTGNNNN) were used to derive five distinct DNA shape features. Integration of these DNA shape feature matrices with sequence features as input for predictive models, help to evaluate the contribution of DNA shape to BMAL1 binding. The mean AUROC scores for the liver, kidney, and heart were 0.97, 0.98, and 0.98, respectively, which are higher than the sequence-only model. Additionally, the mean AUPRC metric increased significantly compared to the sequence-only model, rising from 0.09 to 0.79 for the liver, 0.10 to 0.51 for the kidney, and 0.06 to 0.71 for the heart. This suggests a significant difference in local DNA shape features between the bound and unbound E-boxes. Further analysis revealed that the EP, Roll, and ProT DNA shape features contributed 33% to the prediction of BMAL1 binding to the E-boxes in the liver. In the kidney, the EP, ProT, and MGW DNA shape features contributed 68%, while in the heart, EP, Roll, and MGW contributed 70% to the prediction. Overall, the EP, Roll, MGW, and ProT DNA shape features had the most significant influence on the prediction of bound E-boxes across all three tissues. However, training and evaluating DNA shape-only models yielded lower performance compared to DNA sequence-only models, suggesting that the local shape or configuration of DNA near the E-box alone is insufficient to predict BMAL1 binding. This demonstrates the importance of considering both DNA sequence and shape features in predicting transcription factor binding, with specific DNA shape features playing a crucial role in determining BMAL1 binding affinity to E-boxes across different tissues.

Performance of sequence + shape + histone modification (HM) models: Histone modifications (HMs) within gene promoter and enhancer regions are widely recognized to correlate with the binding of transcription factors (TFs)⁷⁷. Despite this correlation, the intricate mechanisms governing the interaction between TF binding and HMs remain incompletely understood. Recent investigations have shed light on the role of HMs in enhancing the predictive accuracy of models aimed at forecasting TF binding. Notably, the impact of HMs on model performance varies depending on the specific TF under consideration. For instance, models predicting binding of bHLH transcription factors have exhibited significantly enhanced accuracy upon the inclusion of HMs^{78,79}. Based on these findings, several models have been developed to improve TF binding prediction using results from epigenetic assays^{80,81}. We examined the importance of HMs in prediction of BMAL1 binding by augmenting the sequence and DNA shape feature matrix with five histone features: H3K27ac, H3K4me1, H3K4me3, H3K27me3, and H3K36me3. The selection of these HM features was based on both their availability in existing datasets and their established roles in facilitating transcription factor binding⁷⁹. By incorporating these HM features into the models, I obtained mean AUROC scores of 0.99, 0.988, and 0.99 for the liver, kidney, and heart, respectively. Additionally, the mean AUPRC performance increased significantly to 0.95, 0.65, and 0.79 for the liver, kidney, and heart, respectively (Figure 7). These results highlight the importance of considering not only DNA sequence and shape features but also epigenetic information, such as histone modifications, in accurately modeling transcription factor binding events across different tissue types.

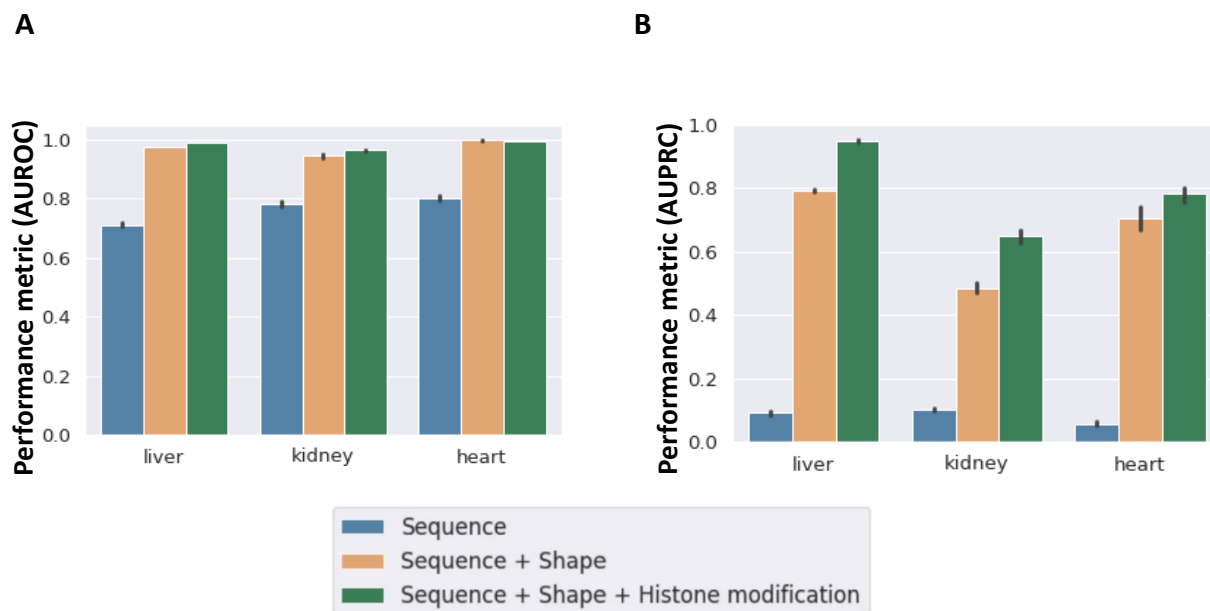


Figure 7: Adding DNA Shape and Histone modification (HM) features to DNA Sequence significantly improves prediction of BMAL1 binding across all tissues. **(A)** The area under the receiver operating characteristics (AUROC) for liver, kidney, and heart for the sequence-only model (blue), sequence plus DNA shape model (brown) and sequence plus DNA shape plus HM model (green). The mean AUROC increases sharply with the addition of DNA shape features to the model, with a much smaller increase associated with the addition of HMs. **(B)** The area under the precision recall curve (AUPRC) in liver, kidney and heart for the sequence-only model (blue), sequence plus DNA shape model (brown) and sequence plus DNA shape plus HM model (green). As with AUROC, the mean AUPRC increased by a large margin with the addition of DNA shape features to the model, with a smaller increase associated with the addition of HMs. Adapted from Marri et al⁷⁵.

	XGBoost		Logistic regression	
	AUROC	AUPRC	AUROC	AUPRC
Liver				
DNA Sequence Only model	0.71 ± 0.00	0.09 ± 0.01	0.71 ± 0.00	0.08 ± 0.01
DNA sequence plus DNA shape model	0.97 ± 0.00	0.79 ± 0.01	0.97 ± 0.00	0.79 ± 0.01
DNA sequence plus histone modification model	0.85 ± 0.02	0.13 ± 0.01	0.80 ± 0.00	0.12 ± 0.03
DNA shape plus histone modification model	0.90 ± 0.01	0.22 ± 0.01	0.81 ± 0.01	0.16 ± 0.01
DNA sequence and shape plus histone modifications model	0.99 ± 0.00	0.95 ± 0.00	0.97 ± 0.00	0.91 ± 0.00
Kidney				
DNA Sequence Only model	0.78 ± 0.01	0.10 ± 0.01	0.77 ± 0.00	0.10 ± 0.01
DNA sequence plus DNA shape model	0.94 ± 0.00	0.50 ± 0.01	0.79 ± 0.01	0.10 ± 0.01
DNA sequence plus histone modification model	0.89 ± 0.00	0.19 ± 0.01	0.87 ± 0.00	0.13 ± 0.01
DNA shape plus histone modification model	0.95 ± 0.01	0.31 ± 0.01	0.88 ± 0.01	0.15 ± 0.01
DNA sequence and shape plus histone modifications model	0.96 ± 0.00	0.65 ± 0.01	0.88 ± 0.01	0.15 ± 0.01
Heart				
DNA Sequence Only model	0.80 ± 0.01	0.06 ± 0.01	0.78 ± 0.01	0.05 ± 0.01
DNA sequence plus DNA shape model	0.99 ± 0.00	0.71 ± 0.03	0.97 ± 0.01	0.49 ± 0.02
DNA sequence plus histone modification model	0.96 ± 0.00	0.26 ± 0.01	0.92 ± 0.01	0.8 ± 0.01
DNA shape plus histone modification model	0.97 ± 0.00	0.35 ± 0.00	0.95 ± 0.00	0.22 ± 0.00
DNA sequence and shape plus histone modifications (Heart)	0.99 ± 0.00	0.80 ± 0.04	0.98 ± 0.01	0.47 ± 0.02

Table 1: Model Performance scores: The performance of models predicting BMAL1-DNA binding status in open chromatin of the liver, kidney, and heart using XGBoost and logistic regression.

Table 1 (cont'd)

Performance of each model is represented as a mean value with a 95% confidence interval around the results from 5-fold cross validation. The highest model performance for each tissue is bolded. Adapted from Marri et al⁷⁵.

Feature importance reveals tissue-specific BMAL1 binding grammar.

Given the improved performance of the models incorporating sequence, shape, and histone modification (HM) features, I applied the ELI5 permutation importance method to discern the most predictive features governing BMAL1-DNA binding⁸². The importance of each DNA shape and histone modification feature was computed by aggregating the importance scores across all bins associated with that specific feature. Furthermore, to facilitate comparative analysis, the feature importance of each nucleotide type at a given position relative to the E-Box motif was normalized against the cumulative feature importance at that nucleotide position. The analysis revealed that the immediate flanking sequences upstream and downstream of the core E-box binding motif played pivotal roles in predicting BMAL1 binding across the liver, heart, and kidney tissues as compared to distal flanking sequences (Fig 8). Previous studies examining the binding specificities of bHLH transcription factors CBf1 and Tye7 in yeast have shown that 2-bp flanking sequences contribute to binding of these transcription factors to the E-box⁵⁵. In our quantitative analysis of the E-box sequence, we did not find the two central base pairs of the CANNTG E-box motif to directly contribute to the model performance, even though BMAL1 has a strong preference for the CG central dinucleotide across all three mouse tissues. The nucleotide G at the second proximal upstream flanking sequence (Seq-2) was a robust predictor of BMAL1-DNA binding in the liver (Fig 8A). Remarkably, this nucleotide accounted for over 50% of the feature weights utilized in predicting BMAL1-DNA binding in the liver. Additionally, other influential features included EP (10%) and H3K27ac (6%). The comprehensive assessment of feature importance highlighted the

significance of various DNA shape and histone modification features in predicting BMAL1 DNA binding in the liver, with the majority exhibiting individual weights exceeding 5%, underscoring their critical roles in this process. Conversely, most DNA sequence features, barring Seq-2, exhibited feature weights below the 5% threshold. In the kidney, H3K27ac emerged as the most influential feature, contributing 21% to the overall feature importance (Fig 8B), followed closely by EP, with a feature importance of 19%. Notably, three histone modifications (H3K27ac, H3K4me3, and H3K4me1) and four DNA shape features (EP, ProT, MGW, and Roll) all demonstrated feature weights surpassing 5%. Similarly, in the heart, H3K27ac and H3K4me3 assumed the highest feature importance (both exceeding 20%), followed by EP (8%). Predominantly, DNA sequence features exhibited weights below 5% in both heart and kidney tissues. Across all three tissues, histone modifications (H3K27ac, H3K4me1, H3K4me3) and DNA shape features (EP, Roll) consistently exhibited high importance scores (Fig 8A-C). Notably, H3K27ac and H3K4me1 emerged as the most influential histone modifications across all tissues, with H3K4me3 and H3K36me3 also making substantial contributions, particularly in the kidney and heart. These results show that the combination of the TF binding motif and its flanking sequence, local shape of DNA, and histone modifications is sufficient to produce predictive models of BMAL1 binding to E-box motifs, especially in the mouse liver. The second upstream flanking nucleotide (Seq-2) stood out as the most crucial feature in the liver, with nucleotides G and C overrepresented in this region. This observation was corroborated by the sequence logo of bound E-box sequences, encompassing 4 base pairs upstream and downstream of the core motif (Fig 8D). Further analysis of bound E-box motifs, along with their upstream and downstream flanking sequences, unveiled a notable enrichment of the nucleotide G at the third position of the 5' flanking region, particularly prevalent among liver-bound E-boxes (1228 out of 3374 bound E-boxes) (Fig

9A-B), a trend not mirrored in bound E-box motifs kidney and heart. The histone modifications H3K27ac (associated with active enhancers and promoters), H3K4me1 (enriched at enhancer regions), and H3K4me3 (found at active promoters) played significant roles in determining BMAL1 binding affinity across different tissues.

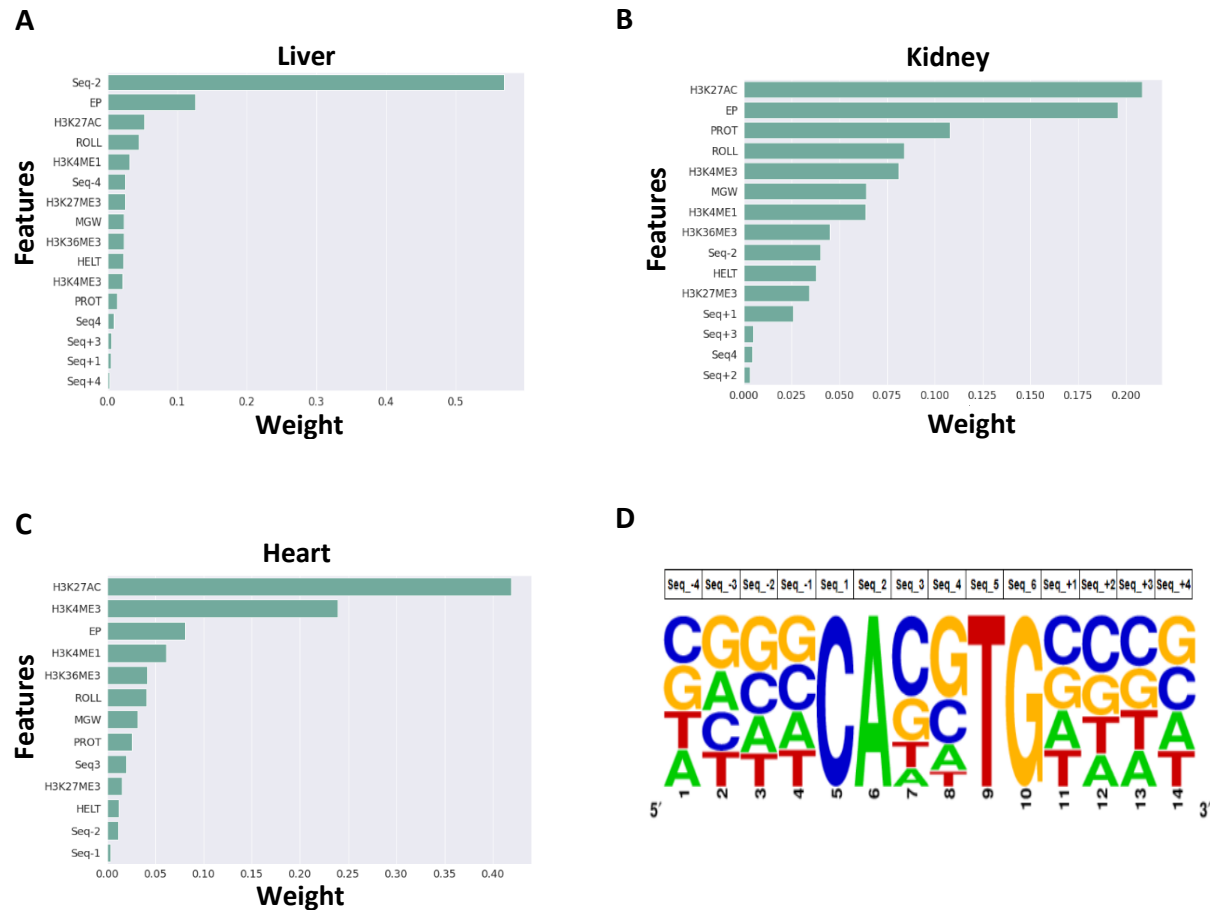
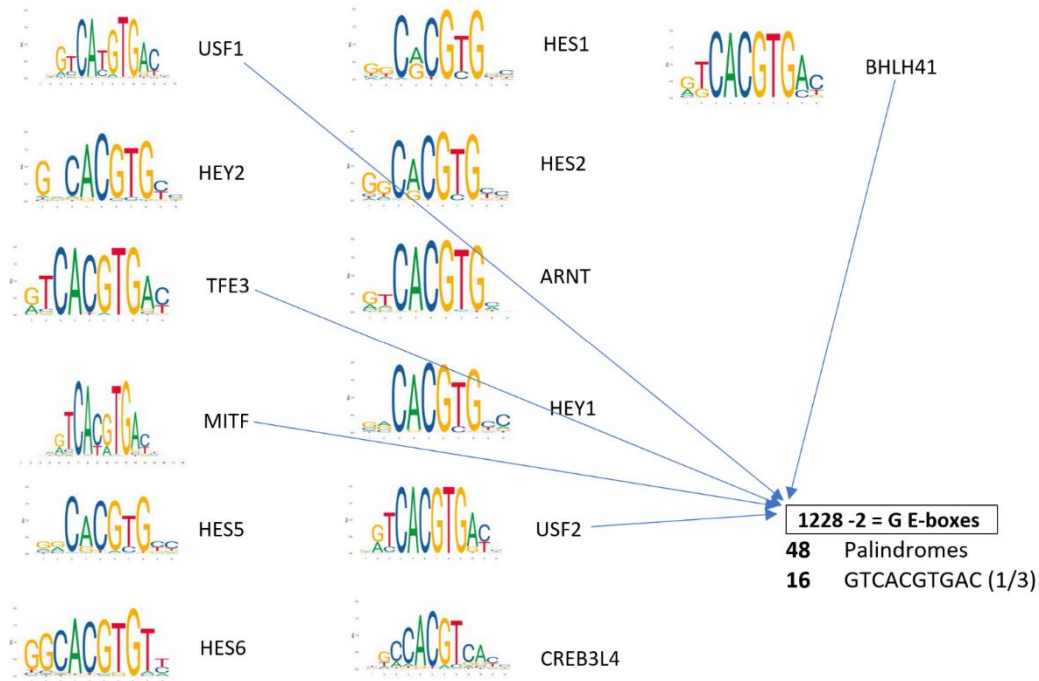


Figure 8: Feature importance of all genomic features (sequence and DNA shape) and epigenomic (histone modification) features from the XGBoost classifier model across all tissues. Feature importance in the XGBoost classifier model in (A) liver, (B) kidney, and (C) heart. The feature importance for each DNA shape and histone modification feature is calculated as the sum of all the feature importance of all bins for that particular histone modification feature. The feature importance of each nucleotide type at a particular position relative to the E-Box motif is normalized to the nucleotide type and the sum of all feature importance at that nucleotide position. (D) Standard plot sequence logo for BMAL1 bound E-box motifs in the liver⁸³. Adapted from Marri et al⁷⁵.

A



B

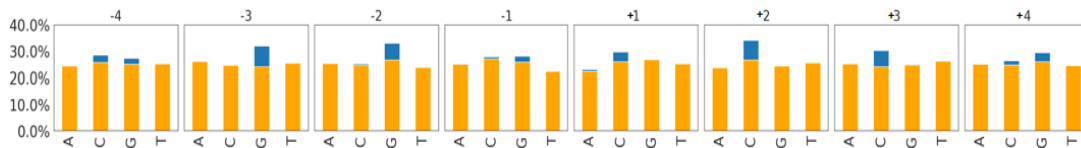


Figure 9: Analysis of Liver bound E-box motifs to investigate the importance of the nucleotide G in the third position of the 5' flanking sequence. (A) Analysis of the bound E-box motif with their upstream and downstream flanking sequence revealed that the nucleotide G at the position third of the 5' flanking sequence is enriched in bound E-box motifs in the liver. 1228 out of 3374 of the motifs have nucleotide G at the third position of the 5' flanking sequence. 48 out of the 1228 are palindromes and 16 out of the 48 are the sequence GTCACGTGAC. (B) Percentage of enriched flanking sequence nucleotide in the liver E-box motifs. (Orange bar corresponds to unbound E-box motifs and blue bar corresponds to bound E-box motifs). Adapted from Marri et al⁷⁵.

Cross-tissue models highlight differences in BMAL1-DNA binding among tissues.

To test the hypothesis that the DNA binding of BMAL1 is determined by similar factors across the three tissues, we formulated cross-tissue models aimed at predicting binding with features based on - (a) sequence only; (b) sequence plus DNA shape; and (c) sequence plus DNA shape plus histone modifications. We trained these models on all data available in the respective tissue, using the optimal hyper-parameters previously derived for the respective within-tissue model. Models trained on one tissue were used to predict BMAL1 binding in a different tissue. The performance of the sequence-only models trained on tissue X and predicting tissue Y (X_Y model) was comparable to the performance of the within-tissue sequence-only model in tissue X, for all tissues (Fig 10A). However, the incorporation of DNA shape and histone modification (HM) features lead to a reduction in performance scores across some cross-tissue models in contrast to the sequence-only models (Fig 10A-C). Notably, the sequence plus shape model, trained on liver data, aptly classified 22% of the E-boxes bound in both kidney (liver_kidney) and heart (liver_heart) (Fig 10C). This model predicted most of the bound E-boxes in the kidney and heart as unbound, yielding a high false negative rate. The inclusion of histone modification features improved AUROC and AUPRC metrics for most cross-tissue models (Fig 10 A-B) . However, the cross-tissue sequence plus DNA shape plus HM model, trained on liver data, correctly classified only 18% of E-boxes bound in the kidney and 19% in the heart, manifesting a higher false negative rate than the sequence plus shape model. The cross-tissue analysis revealed an intriguing pattern when models trained on kidney and heart data were evaluated on liver data (kidney_liver and heart_liver). In these cases, the inclusion of DNA shape and histone modification (HM) features led to a significant improvement in performance compared to other model types. Specifically, the AUROC score of the kidney_liver model increased from 0.68 for the sequence plus DNA shape

model to 0.83 for the sequence plus DNA shape plus HM model. Furthermore, the AUPRC score exhibited a sharp increase, from 0.055 to 0.38, when HM features were added. In contrast to the previous findings, where the addition of DNA shape and HM features generally decreased the performance of cross-tissue models, the kidney_liver and heart_liver models demonstrated a significant improvement in predictive power when these genomic and epigenetic features were incorporated. These results suggest that while sequence-only models performed relatively poorly in cross-tissue binding prediction, the inclusion of additional features, such as DNA shape and histone modifications, can enhance the predictive capability of cross-tissue models in certain cases. However, the effectiveness of these features appears to be tissue-specific, as evidenced by the contrasting results obtained when predicting binding in the liver versus predicting binding in other tissues. Overall, the observed tissue specificity of BMAL1 DNA binding highlights the complexity of the regulatory mechanisms governing this transcription factor's binding patterns. The interplay between sequence features, DNA shape, and epigenetic modifications, such as histone modifications, appears to be context-dependent, with varying degrees of influence across different tissue types. These findings underscore the importance of considering tissue-specific factors in the development of accurate binding prediction models for transcription factors like BMAL1.

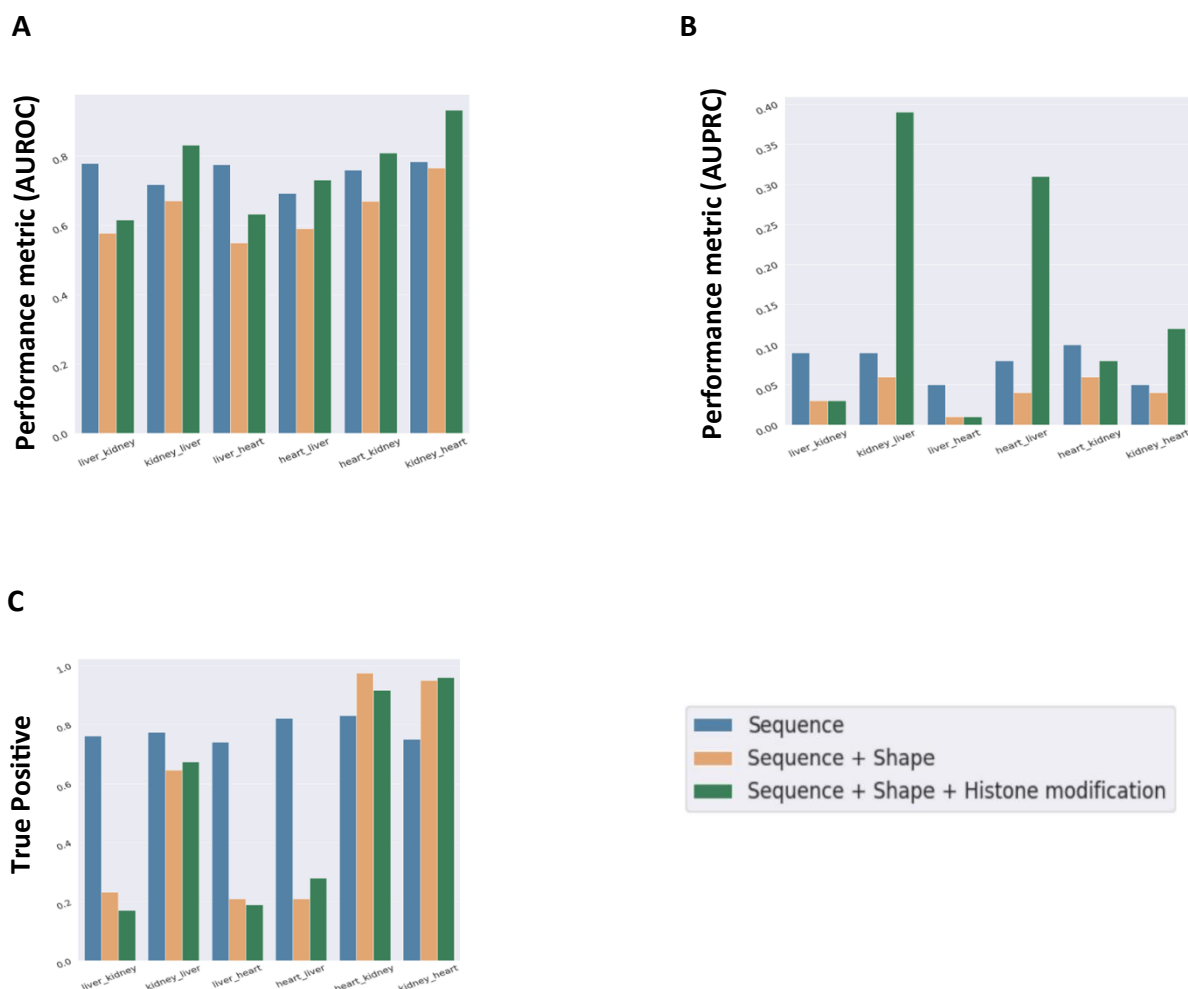


Figure 10: Performance metrics for cross-tissue prediction models. Scores for the liver, kidney and heart sequence-only model (blue bars), sequence plus DNA shape model (brown bars) and sequence plus DNA shape plus HM model (green bars): (A) Area under the receiver operating characteristics (AUROC); (B) Area under the precision recall curve (AUPRC); (C) True positive rates. (Notation explanation: liver_kidney refers to the model trained on the liver dataset and used to predict binding on the kidney dataset). Adapted from Marri et al⁷⁵.

CHAPTER 3: MATHEMATICAL MODELING OF THE SPATIAL AND TEMPORAL DYNAMICS OF CIRCADIAN CLOCK IN THE LIVER

INTRODUCTION

In the previous chapter, the genetic and epigenetic determinants of BMAL1-DNA binding within peripheral tissues were elucidated. In this chapter, the focus will be on investigating the spatiotemporal expression patterns of circadian clock genes within the liver lobule, as well as exploring the factors that contribute to the synchronization of their expression. Within the intricate landscape of the brain, the suprachiasmatic nucleus (SCN) emerges as the central orchestrator, functioning as the master pacemaker of the circadian system. It coordinates and synchronizes cellular, tissue-specific, and systemic rhythms, thereby regulating vital biological processes^{84,85}, including the regulation of body temperature, glucose metabolism, sleep-wake patterns, hormone secretion, and bone formation^{86–88}. Disruption of the circadian cycle by environmental stimuli has been linked to several pathological conditions, including cardiovascular disease, diabetes, bipolar disorder, obesity, and cancer^{39,89–91}. At the molecular level, a complex network of transcriptional, translational, and post-translational feedback loops intricately governs the generation of circadian rhythms, not only within the SCN but also in peripheral tissues such as the liver^{10,92}. At the core of this regulatory network lies a group of transcriptional activators including circadian locomotor output cycles kaput (Clock), Neuronal PAS domain protein 2 (Npas2), brain and muscle ARNT Like 1 (Bmal1), and the retinoic acid-related orphan receptors (Rora, Rorb, Rorc). These activators act in concert with a cohort of repressors, notably the period genes (Per1, Per2, Per3), the cryptochrome genes (Cry1, Cry2), and the reverb-clear orphan receptors (Reverba, Reverbb), to establish the rhythmicity of gene expression^{75,84}.

As discussed in the previous chapter, the master regulator heterodimer composed of CLOCK-BMAL1 (or NPAS2-BMAL1), binds to specific DNA motifs known as E-boxes within the

regulatory regions of target genes such as *Per*, *Cry*, *Ror*, and *Reverb*, thereby promoting their transcriptional activation. Subsequently, the *PER* and *CRY* proteins form a cytoplasmic complex, which translocates into the nucleus to create a negative feedback loop by inhibiting the transcription of their own genes. This intricate interplay also extends to the suppression of *Ror* and *Reverb* gene transcription. Further complexity arises from the competitive binding of *ROR* and *REV-ERB* proteins to the *ROR* regulatory element (RRE) within the promoter region of the *Bmal1* gene. Here, *ROR* acts as an activator, while *REV-ERB* serves as a repressor, thus exerting tight control over the transcriptional regulation of *Bmal1* and consequently influencing the overall circadian rhythm^{92,93}.

The liver stands as a pivotal peripheral oscillator within the intricate machinery of mammalian physiology, serving as a crucial hub for the interplay of anabolic and catabolic processes involving lipids and amino acids⁹⁴. Structurally, the mammalian liver is organized into functional units known as "lobules," each comprising predominantly hepatocytes, the primary cellular constituents of liver parenchyma. Hepatocytes in the lobule are aligned along a distinct axis extending from the portal vein to the central vein, thereby facilitating their categorization based on their proximity to either endpoint (Fig 11). Gene expression and resulting metabolic functions exhibit a spatial gradation along the portal-to-central axis of the lobule. Specifically, processes such as gluconeogenesis and β -oxidation predominate at the portal end, whereas glycolysis and lipogenesis are enriched towards the central region. This spatial organization underscores the liver's intricate orchestration of metabolic pathways to ensure optimal nutrient utilization and energy balance⁹⁵. Recent insights from single-cell gene expression studies have unveiled an intriguing aspect of liver physiology: the expression of core circadian clock genes manifests in a non-zonated manner across hepatocytes along the portal-to-central axis. Unlike metabolic functions, where spatial gradients

are evident, circadian gene expression appears to be uniformly distributed across the lobular axis. This observation hints at a sophisticated mechanism whereby hepatocytes synchronize their circadian rhythms across the entire liver tissue, despite their spatial heterogeneity²⁹. This phenomenon of synchronized circadian oscillations across hepatocytes has been attributed to a coupling mechanism that harmonizes gene expression among cells along the portal-to-central axis, thereby generating coordinated and synchronized temporal rhythms^{96–98}. While the precise molecular machinery orchestrating this synchronization remains elusive, emerging evidence suggests a potential role for transforming growth factor–beta (TGF- β) as a putative coupling factor. However, the intricacies of how TGF- β or other potential mediators precisely regulate the synchronized oscillations in liver gene expression warrant further investigation^{98–100}. Unlike the established neurotransmitter-mediated coupling mechanism in the suprachiasmatic nucleus (SCN) regulating central nervous system circadian rhythms¹⁰¹, the mechanisms behind synchronized liver gene expression oscillations are fully not understood. Understanding this synchronization would offer insights into tissue-level circadian regulation and novel therapeutic strategies for metabolic disorders.

We developed a spatiotemporal multicellular mathematical model of the mammalian liver circadian clock regulatory network^{9,102–104}. Two sets of models were developed for the mouse hepatic clock: 1) Model 1, which assumed no communication among the cellular oscillators, leading to non-synchronized gene expression in hepatocytes across the central to portal axis of the lobule; and 2) Model 2, which incorporated communication among cells, resulting in synchronized gene expression across the lobular axis. The analysis revealed a positive correlation between the amplitudes of the observables (gene expression levels or protein levels) and their respective transcription rate parameters, while a negative correlation was observed between the amplitudes

and their respective degradation rates. This finding highlights the crucial role of transcription and degradation rates in determining the oscillatory amplitudes of the circadian clock components. To validate the models, the transcription and degradation rate parameters were estimated from single-cell RNA-Seq data¹⁰². The simulated results exhibited a high correlation with the experimental data, with an $R^2 > 0.9$, indicating a strong agreement between the model predictions and the observed gene expression patterns. In summary, this study computationally analyzed asynchronous and synchronous spatial and temporal circadian oscillations in the mammalian liver. The models revealed the dependence of the oscillatory amplitudes of circadian clock components on their respective transcription and degradation rates. The high correlation between simulated and experimental data demonstrated the robustness of the mathematical models in capturing the intricate dynamics of the liver circadian clock regulatory network, and the complex interplay between spatial organization, cellular communication, and circadian rhythms.

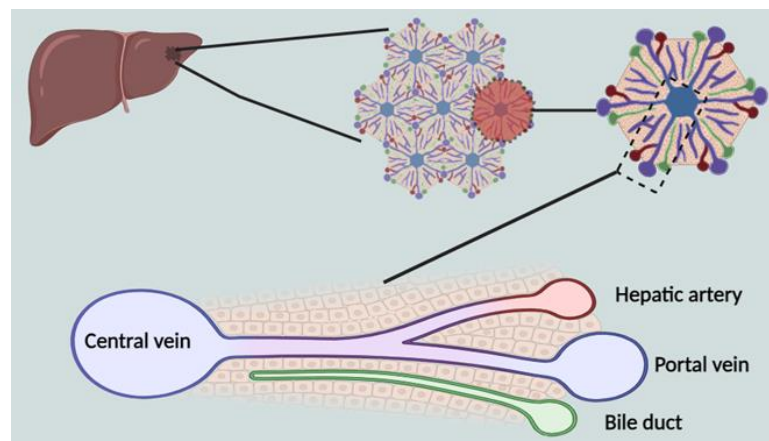


Figure 11: Schematic of a cross-section through the lobule, the fundamental structural unit of the mammalian liver, predominantly made up of hepatocytes extending in layers from the portal triad to the central vein (generated using <https://biorender.com/>).

METHODS

Model design.

This model was developed to study the communication among single cellular circadian oscillators (i.e. hepatocytes) in the liver lobule, or intercellular “coupling”. This coupling ensures a coherent output in the suprachiasmatic nuclei (SCN) and other peripheral tissues. In our model, we assembled the key interactions within the clock gene network from the literature. Five known clock genes were used in developing this model. For simplicity, we do not distinguish between the multiple isoforms in these genes. For example, Per1, Per2 and Per3 are represented by a single Per gene as a model variable. The same reasoning applies to the proteins and respective protein complexes. The gene network used in this model consist of the interactions among (Per, Cry, Ror, Rev-erb and Bmal1). Per, Cry, and Rev-erb act as transcriptional repressors and Ror, and Bmal1 as activators. The interactions among these genes were then transformed to a wiring diagram that consist of two main dependent feedback loops: the Per-Cry negative feedback loop and the Ror, Rev-erb, Bmal1 positive feedback loop.

The central component of the model, the CLOCK_BMAL transcription factor complex binds to the promoter regions of the clock genes (Per, Cry, Ror, Rev-erb) activating their transcription. These various mRNAs are translated into the respective proteins in the cytosol. The RORc and REV-ERBc protein in the cytosol are phosphorylated reversibly and unphosphorylated RORc and REV-ERBc are transported to the nucleus. In the nucleus, RORn binds to the promoter region of the clock gene Bmal1 activating its transcription. The Bmal1 mRNA is translated to BMAL1c protein in the cytosol which is phosphorylated reversibly. Unphosphorylated BMAL1c is transported in the nucleus where it forms a reversible heterodimer with the CLOCK protein denoted as CLOCK_BMAL. The nuclear REV-ERBn protein binds to the promoter region of the

Bmal1 gene to repress the activity of the nuclear RORn protein. The PER and CRY proteins in the cytosol are phosphorylated reversibly. Unphosphorylated PER and CRY protein in the cytosol form the reversible PER-CRY heterodimer which is then transported into the nucleus. PER-CRY represses the transcription of their own genes (Per and Cry) by binding with the CLOCK_BMAL protein. This association represses the transcriptional activities of all the genes that are activated by CLOCK_BMAL. The dissociation of CLOCK_BMAL from the PER-CRY complex allows CLOCK_BMAL to bind to the promoter regions of clock genes (Per, Cry, Ror, Rev-erb) starting the process all over again.

Two sets of models were developed in this study. (1) Model without communication between the cellular oscillators leading to non-synchronized gene expression across the central to portal axis of the liver lobule and (2) model with communication between cells leading to synchronized gene expression across the lobule.

Model 1: Model without communication between the cellular oscillators.

- The first five ordinary differential equations (ODEs) represent the transcription of *Per*, *Cry*, *Bmal1*, *Ror* & *Rev-erb* genes into their respective mRNA's denoted by M_P , M_C , M_B , M_{Ro} & M_{Re} respectively by the nuclear *CLOCK_BMAL* protein (CB_n), *PER_CRY* protein (PC_n), *ROR* protein (Ro_n) and *REV_ERB* protein (Re_n).

$$\frac{dM_P}{dt} = \frac{V_{sp} * (CLOCK_BMAL1)^n}{Kip^n + (CLOCK_BMAL1 * PER_CRY)^m + (CLOCK_BMAL1)^n} - d_p * M_P$$

$$\frac{dM_C}{dt} = \frac{light * V_{sc} * (CLOCK_BMAL1)^o}{Kic^o + (CLOCK_BMAL1 * PER_CRY)^p + (CLOCK_BMAL1)^o} - d_c * M_C$$

$$\frac{dM_B}{dt} = \frac{V_{sb} * (ROR)^p}{Kib^p + (ROR * REV_ERB)^q + (ROR)^p} - d_b * M_B$$

$$\frac{dM_R}{dt} = \frac{V_{sr} * (CLOCK_BMAL1)^r}{Kir^r + (CLOCK_BMAL1 * PER_CRY)^s + (CLOCK_BMAL1)^r} - d_r * M_R$$

$$\frac{dM_{Re}}{dt} = \frac{V_{sre} * (CLOCK_BMAL1)^t}{Kire^t + (CLOCK_BMAL1 * PER_CRY)^u + (CLOCK_BMAL1)^t} - d_{re} * M_{Re}$$

- The next six ODEs represent the translation and reversible activities of the unphosphorylated PER, CRY, BMAL1, REV-ERB, ROR protein and PER-CRY protein complex in the cytoplasm denoted by P_c , C_c , B_c , Re_c , Ro_c & PC_c respectively.

$$\begin{aligned} \frac{dP_c}{dt} = & k * Per_mRNA + Kpc1 * (CYTOSOLIC_PER_CRY_PROTEIN) - Kpc0 * (CYTOSOLIC_PER_PROTEIN) \\ & * (CYTOSOLIC_CRY_PROTEIN) - Kpc * ((CYTOSOLIC_PER_PROTEIN)) + \\ & Kppc * (PHOS_CYTOSOLIC_PER_PROTEIN) - dpc * (CYTOSOLIC_PER_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dC_c}{dt} = & k1 * Cry_mRNA + Kpc1 * (CYTOSOLIC_PER_CRY_PROTEIN) - Kpc0 * (CYTOSOLIC_PER_PROTEIN) \\ & * (CYTOSOLIC_CRY_PROTEIN) - Kcc * ((CYTOSOLIC_CRY_PROTEIN)) \\ & + Kcpc * (PHOS_CYTOSOLIC_CRY_PROTEIN) - dcc * (CYTOSOLIC_CRY_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dB_c}{dt} = & k2 * Bmal1_mRNA - Kbcc * (CYTOSOLIC_BMAL1_PROTEIN) - Kbc * ((CYTOSOLIC_BMAL1_PROTEIN)) \\ & + Kbpc * (PHOS_CYTOSOLIC_BMAL1_PROTEIN) - dbc * (CYTOSOLIC_BMAL1_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dRo_c}{dt} = & k3 * Ror_mRNA - Krcc * (CYTOSOLIC_ROR_PROTEIN) - Krc * ((CYTOSOLIC_ROR_PROTEIN)) \\ & + Krpc * (PHOS_CYTOSOLIC_ROR_PROTEIN) - drc * (CYTOSOLIC_ROR_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dRe_c}{dt} = & k4 * Rev_erb_mRNA - Krecc * (CYTOSOLIC_REV_ERB_PROTEIN) - \\ & Krec * ((CYTOSOLIC_REV_ERB_PROTEIN)) + Krepc * (PHOS_CYTOSOLIC_REV_ERB_PROTEIN) \\ & - drec * (CYTOSOLIC_REV_ERB_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dPC_c}{dt} = & Kpc0 * ((CYTOSOLIC_PER_PROTEIN) * (CYTOSOLIC_CRY_PROTEIN)) \\ & - Kpcc * (CYTOSOLIC_PER_CRY_PROTEIN) - Kpc1 * (CYTOSOLIC_PER_CRY_PROTEIN) \\ & - Kpcp * (CYTOSOLIC_PER_CRY_PROTEIN) + Kpcpc * (PHOS_CYTOSOLIC_PER_CRY_PROTEIN) \\ & - dpcc * (CYTOSOLIC_PER_CRY_PROTEIN) \end{aligned}$$

- The next six ODEs represent the reversible activities of the phosphorylated PER, CRY, BMAL1, REV-ERB, ROR protein and PER-CRY protein complex in the cytoplasm denoted by P_{pc} , C_{pc} , B_{pc} , Re_{pc} , Ro_{pc} & PC_{pc} respectively.

$$\frac{dP_{pc}}{dt} = K_{pc} * (CYTOSOLIC_PER_PROTEIN) - K_{ppc} * (PHOS_CYTOSOLIC_PER_PROTEIN) - d_{ppc} * (PHOS_CYTOSOLIC_PER_PROTEIN)$$

$$\frac{dC_{pc}}{dt} = K_{cc} * (CYTOSOLIC_CRY_PROTEIN) - K_{cpc} * (PHOS_CYTOSOLIC_CRY_PROTEIN) - d_{cpc} * (PHOS_CYTOSOLIC_CRY_PROTEIN)$$

$$\frac{dB_{pc}}{dt} = K_{bc} * (CYTOSOLIC_BMAL1_PROTEIN) - K_{bpc} * (PHOS_CYTOSOLIC_BMAL1_PROTEIN) - d_{bc} * (PHOS_CYTOSOLIC_BMAL1_PROTEIN)$$

$$\frac{dRo_{pc}}{dt} = K_{rc} * ((CYTOSOLIC_ROR_PROTEIN)) - K_{rpc} * (PHOS_CYTOSOLIC_ROR_PROTEIN) - d_{rc} * (PHOS_CYTOSOLIC_ROR_PROTEIN)$$

$$\frac{dRe_{pc}}{dt} = K_{rec} * ((CYTOSOLIC_REV_ERB_PROTEIN)) - K_{rep} * (PHOS_CYTOSOLIC_REV_ERB_PROTEIN) - d_{rec} * (PHOS_CYTOSOLIC_REV_ERB_PROTEIN)$$

$$\frac{dPC_{pc}}{dt} = K_{pcp} * (CYTOSOLIC_PER_CRY_PROTEIN) - K_{pcpc} * (PHOS_CYTOSOLIC_PER_CRY_PROTEIN) - d_{pc} * (PHOS_CYTOSOLIC_PER_CRY_PROTEIN)$$

- The last six ODEs represent the activities of BMAL1, ROR, REV-ERB, CLOCK-BMAL, PER-CRY, and PER-CRY/CLOCK-BMAL proteins in the nucleus denoted by B_n , Ro_n , Re_n , CB_n , PC_n , PC/CB_n , respectively.

$$\frac{dB_n}{dt} = K_{bcc} * (CYTOSOLIC_BMAL1_PROTEIN) - K_{clbn} * (NUCLEAR_BMAL1_PROTEIN) - d_{bn} * (NUCLEAR_BMAL1_PROTEIN)$$

$$\frac{dRo_n}{dt} = K_{rcc} * (CYTOSOLIC_ROR_PROTEIN) - K_{rn} * (NUCLEAR_ROR_PROTEIN) - d_{rn} * (NUCLEAR_ROR_PROTEIN)$$

$$\frac{dRe_n}{dt} = Krecc * (CYTOSOLIC_REV_ERB_PROTEIN) - Kren * (NUCLEAR_REV_ERB_PROTEIN) - dren * (NUCLEAR_REV_ERB_PROTEIN);$$

$$\begin{aligned} \frac{dCB_n}{dt} = & Kclbn * (NUCLEAR_BMAL1_PROTEIN) \\ & - Kcbpc * (NUCLEAR_CLOCK_BMAL1_PROTEIN) * (NUCLEAR_PER_CRY_PROTEIN) \\ & + Kdcbpc * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \\ - dclbn * & (NUCLEAR_CLOCK_BMAL1_PROTEIN) + dpcn * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dPC_n}{dt} = & Kpcc * (CYTOSOLIC_PER_CRY_PROTEIN) \\ & - Kcbpc * (NUCLEAR_CLOCK_BMAL1_PROTEIN) * (NUCLEAR_PER_CRY_PROTEIN) \\ & + Kdcbpc * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \\ - dpcn * & (NUCLEAR_PER_CRY_PROTEIN) + dclbn * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dPC/CB_n}{dt} = & Kcbpc * (NUCLEAR_CLOCK_BMAL1_PROTEIN) * (NUCLEAR_PER_CRY_PROTEIN) \\ & - Kdcbpc * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \\ & - dclbn * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \\ & - dpcn * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \end{aligned}$$

Model 2: Model with communication between the cellular oscillators.

To incorporate the synchronicity of cells in the second model, we assume a global coupling among N cells through cell-to-cell communication. Using a parameter K_f (where K_f is the sensitivity of individual cell oscillator to another cell oscillator, i.e. coupling strength) to describe the communication strength, we calculated a second parameter F, which is the mean communication signal among N cells in the system. This approach was initially used in a three-variable model by Gonze et al¹⁰⁵ We assume the cell-to-cell communication leading to synchronization is mediated by a coupling protein. Previous studies have assumed Transforming growth factor beta (TGF-β) to be the coupling factor responsible for synchronization of the cells to produce a sustained oscillation⁹⁸

- The time-evolution for a single synchronization factor (M , which will later be used to calculate the mean field F) in the cellular medium mediated by a coupling protein is given as:

$$\frac{dM}{dt} = prt * X_i - \frac{vc * M}{dc + M}$$

Where $i = 1...5$ represents each of the observable genes used in the model. X_1 for *Per*, X_2 for *Cry*, X_3 for *Bmal1*, X_4 for *Ror*, and X_5 for *Rev-erb*. There are delays with respect to the clock genes X_i due to transcription, translation, phosphorylation, diffusion, etc. Assuming a linear production equation of the synchronization factor by the clock genes and a nonlinear diffusion rate, prt represents synchronization protein concentration, vc represents Maximum rate of synchronization factor synthesis and dc represents Activation constant for enhancement of synchronization factor synthesis.

The mean field F was then calculated as:

$$F = \frac{1}{N} * \sum_{i=1}^N M_i$$

- The first five ODEs of this model represents the transcription of *Per*, *Cry*, *Bmal1*, *Ror* & *Rev-erb* genes into their respective mRNA's denoted by M_P , M_C , M_B , M_{Ro} & M_{Re} induced by the nuclear proteins *CLOCK_BMAL* (CB_n), *PER_CRY* protein (PC_n), *ROR* protein (Ro_n) and *REV_ERB* (Re_n).

$$\frac{dM_P}{dt} = \frac{V_{sp} * (CLOCK_BMAL1)^n}{K_{ip}^n + (CLOCK_BMAL1 * PER_CRY)^m + (CLOCK_BMAL1)^n} + \frac{V_{dp} * K_f * F}{K_c + K_f * F} - d_p * M_P$$

$$\frac{dM_C}{dt} = \frac{light * V_{sc} * (CLOCK_BMAL1)^o}{K_{ic}^o + (CLOCK_BMAL1 * PER_CRY)^p + (CLOCK_BMAL1)^o} + \frac{V_{dp} * K_f * F}{K_c + K_f * F} - d_c * M_C$$

$$\frac{dM_B}{dt} = \frac{V_{sb} * (ROR)^p}{Kib^p + (ROR * REV_ERB)^q + (ROR)^p} + \frac{Vdp * K_f * F}{K_c + K_f * F} - d_b * M_B$$

$$\frac{dM_R}{dt} = \frac{V_{sr} * (CLOCK_BMAL1)^r}{Kir^r + (CLOCK_BMAL1 * PER_CRY)^s + (CLOCK_BMAL1)^r} + \frac{Vdp * K_f * F}{K_c + K_f * F} - d_r * M_R$$

$$\frac{dM_{Re}}{dt} = \frac{V_{sre} * (CLOCK_BMAL1)^t}{Kire^t + (CLOCK_BMAL1 * PER_CRY)^u + (CLOCK_BMAL1)^t} + \frac{Vdp * K_f * F}{K_c + K_f * F} - d_{re} * M_{Re}$$

- The next six ODEs represent the translation and reversible activities of the unphosphorylated PER, CRY, BMAL1, REV-ERB, ROR protein and PER-CRY protein complex in the cytoplasm denoted by P_c , C_c , B_c , Re_c , Ro_c & PC_c respectively

$$\begin{aligned} \frac{dP_c}{dt} = & k * Per_mRNA + Kpc1 * (CYTOSOLIC_PER_CRY_PROTEIN) - Kpco * (CYTOSOLIC_PER_PROTEIN) \\ & * (CYTOSOLIC_CRY_PROTEIN) - Kpc * ((CYTOSOLIC_PER_PROTEIN)) + \\ & Kppc * (PHOS_CYTOSOLIC_PER_PROTEIN) - dpc * (CYTOSOLIC_PER_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dC_c}{dt} = & k1 * Cry_mRNA + Kpc1 * (CYTOSOLIC_PER_CRY_PROTEIN) - Kpco * (CYTOSOLIC_PER_PROTEIN) \\ & * (CYTOSOLIC_CRY_PROTEIN) - Kcc * ((CYTOSOLIC_CRY_PROTEIN)) \\ & + Kcpc * (PHOS_CYTOSOLIC_CRY_PROTEIN) - dcc * (CYTOSOLIC_CRY_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dB_c}{dt} = & k2 * Bmal1_mRNA - Kbcc * (CYTOSOLIC_BMAL1_PROTEIN) - Kbc * ((CYTOSOLIC_BMAL1_PROTEIN)) \\ & + Kbpc * (PHOS_CYTOSOLIC_BMAL1_PROTEIN) - dbc * (CYTOSOLIC_BMAL1_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dRo_c}{dt} = & k3 * Ror_mRNA - Krcc * (CYTOSOLIC_ROR_PROTEIN) - Krc * ((CYTOSOLIC_ROR_PROTEIN)) \\ & + Krpc * (PHOS_CYTOSOLIC_ROR_PROTEIN) - drc * (CYTOSOLIC_ROR_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dRe_c}{dt} = & k4 * Rev_erb_mRNA - Krecc * (CYTOSOLIC_REV_ERB_PROTEIN) - \\ & Krec * ((CYTOSOLIC_REV_ERB_PROTEIN)) + Krepc * (PHOS_CYTOSOLIC_REV_ERB_PROTEIN) \\ & - drec * (CYTOSOLIC_REV_ERB_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dPC_c}{dt} = & K_{pco} * ((CYTOSOLIC_PER_PROTEIN) * (CYTOSOLIC_CRY_PROTEIN)) \\ & - K_{pcc} * (CYTOSOLIC_PER_CRY_PROTEIN) - K_{pc1} * (CYTOSOLIC_PER_CRY_PROTEIN) \\ & - K_{pcp} * (CYTOSOLIC_PER_CRY_PROTEIN) + K_{pcpc} * (PHOS_CYTOSOLIC_PER_CRY_PROTEIN) \\ & - dpcc * (CYTOSOLIC_PER_CRY_PROTEIN) \end{aligned}$$

- The next six ODEs represent the reversible activities of the phosphorylated PER, CRY, BMAL1, REV-ERB, ROR protein and PER-CRY protein complex in the cytoplasm denoted by P_{pc} , C_{pc} , B_{pc} , Re_{pc} , Ro_{pc} & PC_{pc} respectively.

$$\begin{aligned} \frac{dP_{pc}}{dt} = & K_{pc} * (CYTOSOLIC_PER_PROTEIN) - K_{ppc} * (PHOS_CYTOSOLIC_PER_PROTEIN) \\ & - dp_{pc} * (PHOS_CYTOSOLIC_PER_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dC_{pc}}{dt} = & K_{cc} * (CYTOSOLIC_CRY_PROTEIN) - K_{cpc} * (PHOS_CYTOSOLIC_CRY_PROTEIN) \\ & - dc_{pc} * (PHOS_CYTOSOLIC_CRY_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dB_{pc}}{dt} = & K_{bc} * (CYTOSOLIC_BMAL1_PROTEIN) - K_{bpc} * (PHOS_CYTOSOLIC_BMAL1_PROTEIN) \\ & - dbc * (PHOS_CYTOSOLIC_BMAL1_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dRo_{pc}}{dt} = & K_{rc} * ((CYTOSOLIC_ROR_PROTEIN)) - K_{rpc} * (PHOS_CYTOSOLIC_ROR_PROTEIN) \\ & - drc * (PHOS_CYTOSOLIC_ROR_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dRe_{pc}}{dt} = & K_{rec} * ((CYTOSOLIC_REV_ERB_PROTEIN)) - K_{repc} * (PHOS_CYTOSOLIC_REV_ERB_PROTEIN) \\ & - drec * (PHOS_CYTOSOLIC_REV_ERB_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dPC_{pc}}{dt} = & K_{pcp} * (CYTOSOLIC_PER_CRY_PROTEIN) - K_{pcpc} * (PHOS_CYTOSOLIC_PER_CRY_PROTEIN) \\ & - dpcc * (PHOS_CYTOSOLIC_PER_CRY_PROTEIN) \end{aligned}$$

- The last six ODEs represent the activities of BMAL1, ROR, REV-ERB, CLOCK-BMAL, PER-CRY, and PER-CRY/CLOCK-BMAL protein in the nucleus denoted by B_n , Ro_n , Re_n , CB_n , PC_n , PC/CB_n , respectively.

$$\frac{dB_n}{dt} = Kb_{cc} * (CYTOSOLIC_BMAL1_PROTEIN) - Kcl_{bn} * (NUCLEAR_BMAL1_PROTEIN) - dbn * (NUCLEAR_BMAL1_PROTEIN)$$

$$\frac{dRo_n}{dt} = Kr_{cc} * (CYTOSOLIC_ROR_PROTEIN) - Krn * (NUCLEAR_ROR_PROTEIN) - drn * (NUCLEAR_ROR_PROTEIN)$$

$$\frac{dRe_n}{dt} = Kre_{cc} * (CYTOSOLIC_REV_ERB_PROTEIN) - Kren * (NUCLEAR_REV_ERB_PROTEIN) - dren * (NUCLEAR_REV_ERB_PROTEIN);$$

$$\begin{aligned} \frac{dCB_n}{dt} = & Kcl_{bn} * (NUCLEAR_BMAL1_PROTEIN) \\ & - Kcb_{pc} * (NUCLEAR_CLOCK_BMAL1_PROTEIN) * (NUCLEAR_PER_CRY_PROTEIN) \\ & + Kdcb_{pc} * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \\ & - dcl_{bn} * (NUCLEAR_CLOCK_BMAL1_PROTEIN) + dpc_n * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dPC_n}{dt} = & Kp_{cc} * (CYTOSOLIC_PER_CRY_PROTEIN) \\ & - Kcb_{pc} * (NUCLEAR_CLOCK_BMAL1_PROTEIN) * (NUCLEAR_PER_CRY_PROTEIN) \\ & + Kdcb_{pc} * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \\ & - dpc_n * (NUCLEAR_PER_CRY_PROTEIN) + dcl_{bn} * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \end{aligned}$$

$$\begin{aligned} \frac{dPC/CB_n}{dt} = & Kcb_{pc} * (NUCLEAR_CLOCK_BMAL1_PROTEIN) * (NUCLEAR_PER_CRY_PROTEIN) \\ & - Kdcb_{pc} * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \\ & - dcl_{bn} * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \\ & - dpc_n * (NUCLEAR_CLOCK_BMAL1_PER_CRY_PROTEIN) \end{aligned}$$

In the proposed model, various parameters were incorporated to capture the dynamics of the system under investigation. These parameters, along with their definitions and values, are presented in the table below. To account for potential variations and uncertainties, three distinct sets of parameter values were defined, with parameter set 3 serving as the nominal set (wild type parameter; WT) employed in the model simulations. The simulation from parameter set 3 revealed the experimental peak expression for the circadian clock genes in the model. Also, parameter estimation analysis from figure 8 was used to confirm the accurate reflection of parameter set 3 (WT) to true biology. The table lists the parameters, their descriptions, and the corresponding values for each of the three parameter sets.

Parameter	Definition	Set 1	Set 2	Set 2
Vsp	Maximum rate of Per synthesis	2.2	1.7	2.9
Kip	Activation constant for enhancement of Per by nuclear CLOCK/BMAL	0.24	0.302	0.28
d_p	Maximum rate of Per degradation	0.34	0.375	0.125
Vsc	Maximum rate of Cry synthesis	2.3	1.6	2.8
Kic	Activation constant for enhancement of Cry by nuclear CLOCK/BMAL	0.262	0.32	0.17
d_c	Maximum rate of Cry degradation	0.324	0.39	0.44
Vsb	Maximum rate of Bmal1 synthesis	2.32	1.2	3.4
Kib	Activation constant for enhancement of Cry by nuclear Ror	0.13	0.16	0.16
d_b	Maximum rate of Bmal1 degradation	0.46	0.43	0.49
Vsr	Maximum rate of Ror synthesis	2.04	1.6	3.1
Kir	Activation constant for enhancement of Ror by nuclear CLOCK/BMAL	0.27	0.39	0.3
d_r	Maximum rate of Ror degradation	0.372	0.34	0.27
Vsre	Maximum rate of Rev-erb synthesis	2.15	1.8	2.425
Kic	Activation constant for enhancement of Rev-erb by nuclear CLOCK/BMAL	0.24	0.37	0.25
d_re	Maximum rate of Rev-erb degradation	0.382	0.32	0.285
k	Maximum rate of Per mRNA translation	0.408	0.453	0.412
Kpc1	Dissociation rate of PER-CRY protein to PER protein	0.362	0.383	0.34
Kpco	Association rate of PER & CRY protein to PER-CRY protein	0.3	0.484	0.31
Kpc	Rate of PER phosphorylation	0.304	0.204	0.33
Kppe	Rate of PER de-phosphorylation	0.39	0.306	0.32
dpc	Degradation Rate of PER Protein	0.23	0.024	0.10
k1	Maximum rate of Cry mRNA translation	0.39	0.53	0.42
Kcc	Rate of CRY phosphorylation	0.365	0.242	0.4
Kcpc	Rate of CRY de-phosphorylation	0.306	0.31	0.35
dpc	Degradation Rate of PER Protein	0.18	0.023	0.08
k2	Maximum rate of Bmal1 mRNA translation	0.47	0.34	0.53
Kbcc	Rate of cytosol BMAL Protein transported into the nucleus	0.34	0.342	0.48
Kbc	Rate of BMAL phosphorylation	0.30	0.28	0.3
Kbpc	Rate of BMAL de-phosphorylation	0.38	0.39	0.25
dbc	Degradation Rate of BMAL Protein	0.07	0.05	0.07
k3	Maximum rate of Ror mRNA translation	0.42	0.44	0.5
Krcc	Rate of cytosol ROR Protein transported into the nucleus	0.4	0.35	0.504
Krc	Rate of ROR phosphorylation	0.33	0.36	0.4
Krpe	Rate of ROR de-phosphorylation	0.2	0.40	0.16
drc	Degradation Rate of ROR Protein	0.05	0.05	0.05
k4	Maximum rate of Rev-erb mRNA translation	0.42	0.36	0.4
Krecc	Rate of cytosol REV-ERB Protein transported into the nucleus	0.4	0.304	0.48
Krec	Rate of REV-ERB phosphorylation	0.33	0.36	0.36
Krepc	Rate of REV-ERB de-phosphorylation	0.2	0.34	0.33
drec	Degradation Rate of REV-ERB Protein	0.05	0.06	0.05
Kpcc	Rate of cytosol PER-CRY Protein transported into the nucleus	0.37	0.345	0.4
Kpcp	Rate of PER-CRY phosphorylation	0.25	0.24	0.28
Kpcpc	Rate of PER-CRY de-phosphorylation	0.36	0.36	0.39
dpc	Degradation Rate of PER-CRY Protein	0.017	0.017	0.017
dppc	Degradation Rate of Phosphorylated PER Protein	0.023	0.023	0.23
dcpc	Degradation Rate of Phosphorylated CRY Protein	0.020	0.019	0.020
dbpc	Degradation Rate of Phosphorylated BMAL Protein	0.013	0.013	0.015
drpc	Degradation Rate of Phosphorylated ROR Protein	0.02	0.02	0.02
drep	Degradation Rate of Phosphorylated REV-ERB Protein	0.023	0.23	0.02
dpcpc	Degradation Rate of Phosphorylated PER-CRY Protein	0.025	0.025	0.03
Kelbn	Association rate of CLOCK and BMAL protein to form CLOCK-BMAL protein	0.37	0.32	0.47
dbn	Degradation Rate of nuclear BMAL Protein	0.09	0.03	0.05
Krn	Association rate of ROR and REV-ERB protein	0.35	0.31	0.25
drn	Degradation Rate of nuclear ROR Protein	0.07	0.02	0.07
Kren	Association rate of REV-ERB and ROR protein	0.35	0.31	0.35
drn	Degradation Rate of nuclear REV-ERB Protein	0.06	0.02	0.06
Kcbpc	Association rate of CLOCK-BMAL & PER-CRY proteins in the nucleus	0.49	0.25	0.4
delbn	Degradation Rate of nuclear CLOCK-BMAL Protein	0.30	0.15	0.30
dpen	Degradation Rate of nuclear PER-CRY Protein	0.15	0.075	15
prt	synchronization protein concentration	1.34	0.945	1.28
ve	Maximum rate of synchronization factor synthesis	1.734	1.945	1.58
dc	Activation constant for enhancement of synchronization factor synthesisL	0.134	0.145	0.28

Table 2: Model parameters, definition and values.

RESULTS

Circadian clock mechanism – model design.

The development of the model involved identifying and compiling the key regulatory interactions in the mammalian circadian clock gene network from scientific literature^{9,103}. The model postulates a reversible phosphorylation mechanism for all cytosolic circadian clock proteins. This inclusion is supported by the pivotal role phosphorylation plays in orchestrating circadian regulation. The reversible nature of phosphorylation allows for dynamic fluctuations in protein activity, thus contributing to the intricate regulation of circadian processes. This hypothesis is based on evidence that phosphorylation of key clock proteins (e.g., PER, CRY, BMAL1) plays a regulatory role in the circadian system by altering their stability, localization, protein-protein interactions, and transcriptional activity over the circadian cycle. The reversibility of phosphorylation allows for dynamic changes in protein activity and regulation of circadian processes^{104,106}. The ROR_c and REV-ERB_c proteins (where the subscript 'c' stands for 'cytosolic') undergo reversible phosphorylation, after which they are transported to the nucleus. In the nucleus, ROR_n and REV-ERB_n (the subscript 'n' stands for 'nuclear') bind to the promoter region of the Bmal1 gene to regulate its transcription. Specifically, ROR_n acts as an activator, while REV-ERB_n serves as a repressor in this context. Similarly, the Bmal1 gene is translated into the BMAL1_c protein in the cytosol, which subsequently undergoes reversible phosphorylation. BMAL1_c is then transported to the nucleus, where it forms a reversible heterodimer with the CLOCK protein, generating a protein complex termed "CLOCK/BMAL1". In parallel, the PER and CRY proteins undergo reversible phosphorylation within the cytosol. Upon dephosphorylation, the unphosphorylated PER and CRY proteins assemble into the reversible PER-CRY_c heterodimer, which then undergoes nuclear translocation. Within the nucleus, the PER-CRY_n complex exerts its regulatory role by

repressing the transcription of key clock genes, including Per, Cry, Rev-erb, and Ror. This repression is achieved by binding to the CLOCK_BMAL1 complex, thereby dampening its activity. The dissociation of the transcription factor PER-CRY from the promoters of these genes permits CLOCK_BMAL1 to activate the transcription of Per, Cry, Ror, and Rev-erb, thus initiating a new cycle. This cyclical regulatory process forms the basis of the circadian rhythm. These interactions are summarized in a wiring diagram consisting of two main coupled feedback loops: a positive loop involving the activation of Bmal1 transcription by ROR_n, and a negative loop involving the repression of Per, Cry, Ror, and Rev-erb transcription by PER-CRY_n (Fig 12).

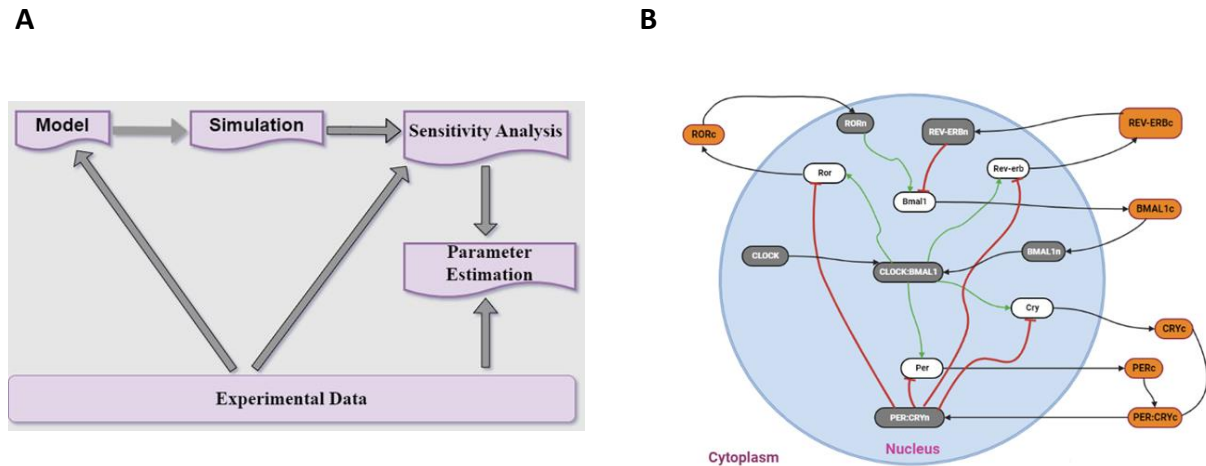


Figure 12: Modeling framework and schematic network diagram. (A) Approach to construction of a deterministic model of the mammalian circadian clock, and data-driven sensitivity analysis and parameter estimation. (B) Schematic diagram of the principal positive and negative regulatory feedback loops in the mammalian clock network. The core negative feedback loop is formed by PER:CRY heterodimers that repress their own transcription. Another regulatory (positive and negative feedback) loop is formed by Ror and Rev-erb competing for binding to ROR/REV-ERB-response element (RORE) to regulate Bmal1. Green lines indicate activation by the transcription factors CLOCK-BMAL1 and ROR_n, red lines indicate repression of genes by transcription factors, and black lines indicate the translation and translocation of mRNAs and proteins. White ovals represent genes, gray ovals proteins in the nucleus and orange ovals proteins in the cytosol.

Multicellular spatiotemporal model of the core clock genes describes coupling among cells.

Gene regulatory networks can be mathematically described by a set of coupled ordinary differential equations (ODEs), where molecular interactions, such as transcription factor binding and gene regulation, are represented by nonlinear functions like Michaelis-Menten or higher-order Hill functions. These nonlinear functions capture the cooperative or saturable nature of these interactions, providing a more accurate representation of the underlying biological processes¹⁰⁷. In this study, the gene regulatory network depicted in Fig 12 B was translated into two deterministic spatiotemporal models, as detailed in the Method section. One model operates in isolation, devoid of intercellular communication (coupling), while the other incorporates intercellular coupling mechanisms. In both models, I employed Michaelis-Menten and Hill function approximations to describe the interactions between transcription factors and genes. Additionally, I assumed mass action and linear kinetics to describe processes such as degradation, translation, and complex formation. These mathematical representations provide a robust framework for simulating the dynamic behavior of the circadian clock gene network. In the absence of synchronization signals that enable cell-cell coupling, autonomous cells in the suprachiasmatic nucleus or other peripheral tissues like the liver would oscillate with different periods, amplitudes, and phases. This desynchronization is due to intrinsic noise and variability in the gene regulatory networks of individual cells. However, intercellular communication induced by extrinsic or intrinsic signals ensures synchrony in the oscillations and resulting biological functions across the cell population. This coupling mechanism, mediated by signaling molecules or direct cell-cell interactions, promotes coordination and coherence in the circadian rhythms and associated processes^{98,103,105}. The two models developed in this study were specifically designed to investigate the spatiotemporal expression patterns of circadian clock genes in the liver. By incorporating

intercellular communication in one of the models, I aimed to explore the role of coupling mechanisms in synchronizing the circadian rhythms across hepatocytes (liver cells). I used a comprehensive dataset comprised of single-cell RNA sequencing (scRNA-seq) and single molecule fluorescence in situ hybridization (smFISH) data, which previously demonstrated that the expression of core circadian clock genes in the liver lobule is non-zonated due to coupling among autonomous cell oscillators²⁹. To dissect the regulatory intricacies underlying these observations, I constructed an uncoupled model by translating the wiring diagram depicted in Fig 12 B into a system of 23 differential equations. These equations encapsulated the dynamic processes of gene transcription and translation, protein complex formation, phosphorylation, and inhibition, thus providing a mechanistic framework to explore the underlying regulatory dynamics. This model was applied to an assembly of over 435 cells spatially arranged to mimic the geometry of a liver lobule (the lobule geometry was described in a piff file for use in the CompuCell3D simulations). To capture potential heterogeneity across cells in transcription rates, which could arise from desynchronization in gene expression, I assumed a Gaussian distribution to randomize the transcription rate parameters for every gene in each individual cell. This stochastic modeling approach aimed to incorporate the inherent variability and noise present in gene expression dynamics across the cellular population within the liver lobule microenvironment. The spatial and temporal expression patterns for each gene were then simulated in the CompuCell3D modeling environment, a powerful platform for simulating and visualizing multicellular systems and their dynamics^{108–110}.

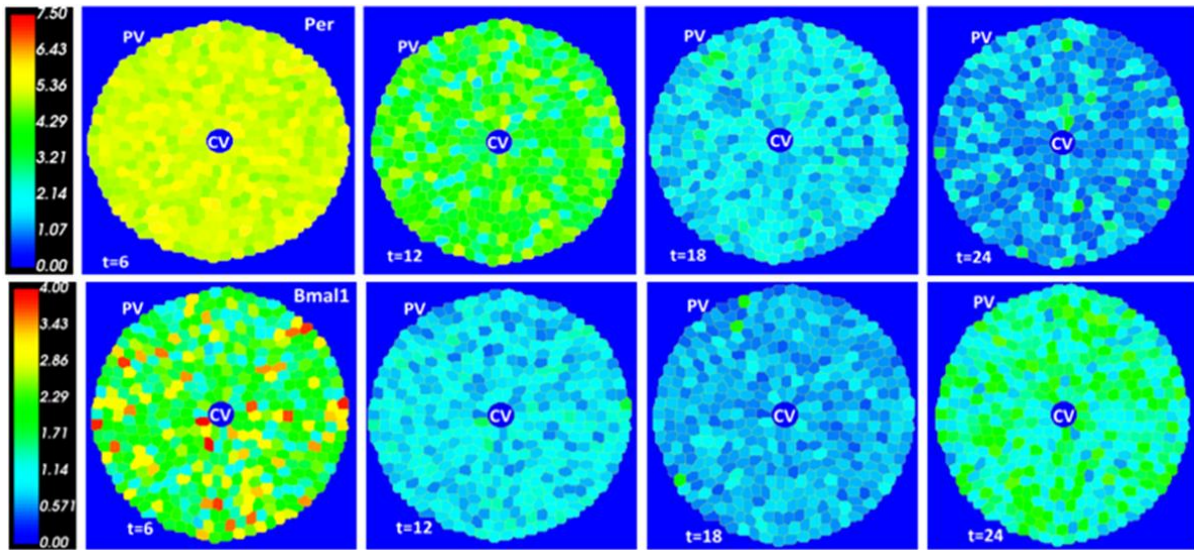
The simulations of the uncoupled model revealed that the mean expression of both *Per* and *Bmal1* genes at timepoints 6, 12, 18, and 24 hours was not zonated across the portal-central axis of the liver lobule (Fig 13 A). This finding confirms results from previous studies, which observed a non-

zonated expression pattern for core circadian clock genes in the liver lobule¹¹¹. In the uncoupled model, the temporal profiles of Per and Bmal1 gene expression exhibited distinctive autonomous oscillations spanning the lobule, characterized by periods ranging from 21 to 28 hours and amplitudes varying between 2.8 to 4.7 (arbitrary units, a.u) for Per, and periods ranging from 21 to 28 hours and amplitudes spanning from 0.8 to 3.2 a.u for Bmal1 (Fig 13 B, E and F). These autonomous oscillations reflect the desynchronization and variability among individual cells in the absence of coupling mechanisms.

To explore the impact of intercellular communication on circadian synchronization, in the coupled model, I assume that synchronization of autonomous cells across the liver lobule is achieved through communication mediated by a putative coupling ligand between each cell and its neighbors. The transmission of this coupling ligand across the liver lobule effectively synchronizes the period of the oscillations. I hypothesized that global synchronization across the lobule is achieved through the average concentration of the coupling ligand affecting clock genes via receptor molecules, as previously suggested^{98,105}. Data from Finger et al⁹⁸. revealed a potential coupling factor: the transforming growth factor-beta (TGF- β), which activates early transcription factors to control the molecular clock machinery, leading to a significant upregulation of Per2 mRNA levels after 2-4 hours. This finding supports the role of TGF- β as a potential coupling ligand in synchronizing the circadian clock across the liver lobule^{103,112}. The spatial and temporal expression for each gene in the model was then simulated in CompuCell3d, showing coupling across the portal-central axis (Fig 13 C). The temporal profiles of Per and Bmal1 gene expression exhibited synchronized oscillations, characterized by a period of 24-25 hours and an amplitude of 5.2 for Per, and a period of 23-25 hours with an amplitude of 2.2 for Bmal1 (Fig 13 D, E, and F). These results demonstrate the impact of intercellular communication and coupling mechanisms on

synchronization of circadian rhythms across the liver lobule. While the uncoupled model showed desynchronized oscillations with varying periods and amplitudes, the coupled model exhibited synchronized oscillations with consistent periods and amplitudes, reflecting the role of coupling factors in coordinating the circadian clock across the cell population.

A



B

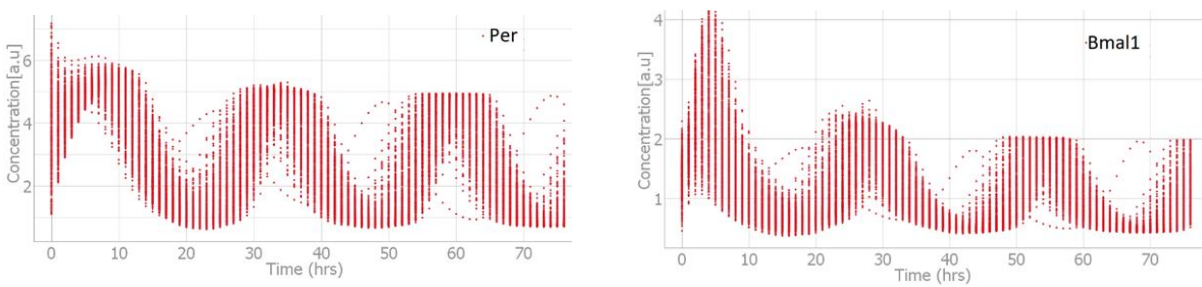
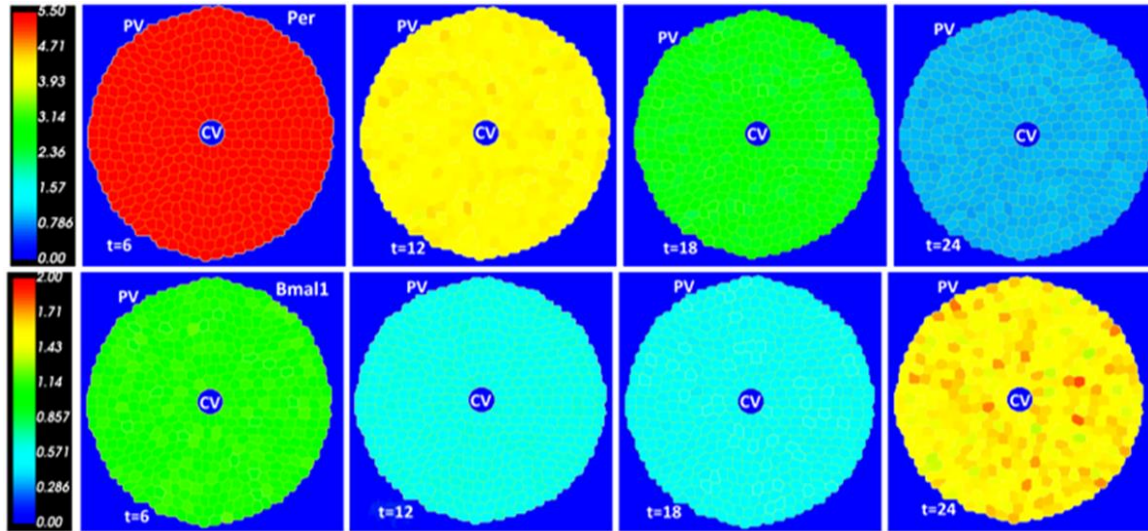


Figure 13: Multicellular spatiotemporal model simulations. Expression across the liver lobule of (A) Period gene (Per) and Brain and Muscle ARNT-Like 1 (Bmal1) without coupling at circadian times ZT 6, 12, 18, and 24 hours. The central- portal lobule display a non-zonated circadian clock genes expression as reported in Droin et al²⁹. (B) Limit cycle oscillations of the Per and Bmal1 genes for 435 cells without coupling with varying period and amplitude values for Per (period: 21-28 hours, amplitude: 2.8-4.7) and Bmal1 (period: 21-28 hours, amplitude: 0.8-3.2).

Figure 13 (cont'd)

(C) The spatial expression of *Per*, and *Bmal1* genes for 435 cells with coupling at circadian time ZT 6 12, 18 and 24 hours across the liver lobule. The expression profile shows a synchrony across all time points. (D) Limit cycle oscillations of the *Per* and *Bmal1* genes with coupling producing synchronized period and amplitude for *Per* (period: 24-25 hours, amplitude: 5.0-5.2) and *Bmal1* (period: 23-25 hours, amplitude: 2.0-2.2). The color bar shows the expression values of genes in each cell across the central (CV) and portal (PV) in the liver lobule. Each dot represents the time-dependent expression of a gene per cell. (E) The distribution of uncoupled *Per* and *Bmal1* oscillation periods for over 400 cells in the model lobule. The period ranges from 21 hours to 28 hours. (F) The distribution of coupled *Per* and *Bmal1* oscillation periods for over 400 cells in the model. The period ranges between 24 hours to 25 hours for *Per* oscillations and 23 hours to 25 hours for *Bmal1* oscillations.

C



D

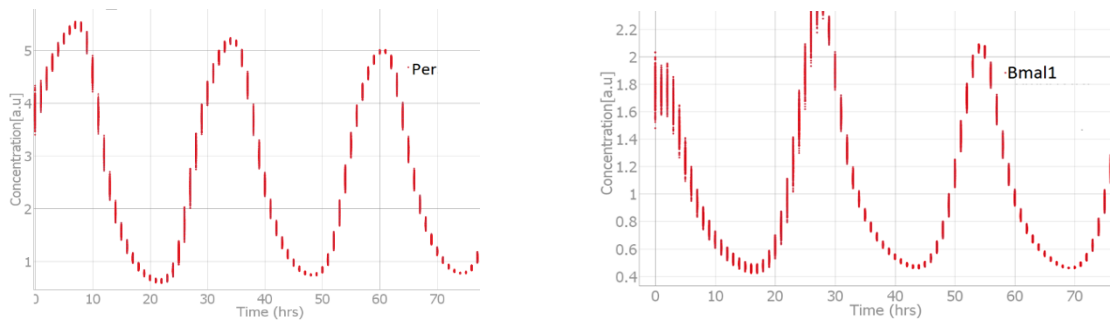
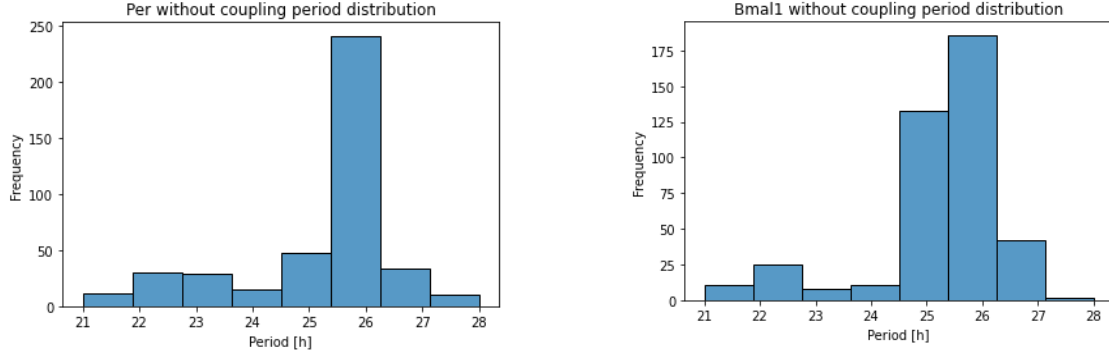
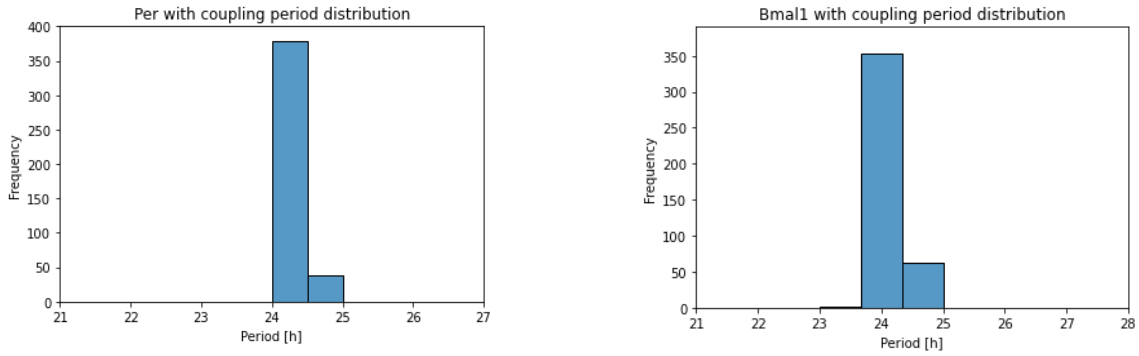


Figure 13 (cont'd)

E



F



Sensitivity analysis

To investigate the sensitivity of the model dynamics to variations in key parameters, I performed a parameter sensitivity analysis using the AMIGO2 toolbox in MATLAB^{113,114}. The model dynamics analyzed included the period of oscillation, amplitude of oscillation, phase shift, and the average amount of mRNA produced. Specifically, the sensitivity analysis focused on the effects of transcription and degradation rate parameters on these model dynamics. I used the Latin hypercube sampling (LHS) method, which is known for providing more precise estimates compared to

random sampling methods¹¹⁵. The LHS technique is used to divide the tested range of each variable into intervals, from which sample points are selected to ensure an efficient exploration of the parameter space.

The LHS method involves the following steps:

- Defining the ranges of the parameters to be tested: For each transcription and degradation rate parameter, I specified a biologically relevant range within which the parameter can vary. The ranges for each parameter were determined from literature values.
- Dividing the parameter ranges into intervals: The LHS method divides the range of each parameter into equal intervals, ensuring that the entire range is adequately sampled.
- Sampling from the intervals: The LHS algorithm selects sample points from each interval in a stratified manner, ensuring that the entire parameter space is covered without clustering or gaps.
- Running simulations: For each set of sampled parameter values, I ran the model simulations and recorded the respective values of the model dynamics (period, amplitude, phase, and average mRNA levels).
- Sensitivity analysis: The sensitivity of each model dynamic to variations in the transcription and degradation rate parameters was quantified by analyzing the relationship between the sampled parameter values and the corresponding model dynamics. Various statistical measures, such as partial rank correlation coefficients (PRCC) or standardized regression coefficients, can be used to assess the sensitivity.

By employing the LHS method and systematically exploring the parameter space, I aimed to identify the most influential parameters affecting the circadian clock dynamics and quantify their

relative impacts. This sensitivity analysis provided valuable insights into the robustness and critical control points of the model, as well as potential targets for experimental validation or therapeutic interventions.

The transcription and degradation rate parameters of the *Reverb* gene respectively show negative and positive correlation with the average expression level of *Bmal1*.

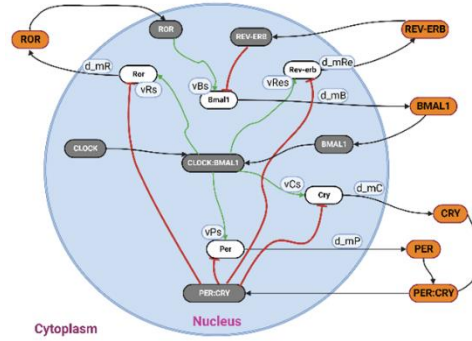
I investigated the effect of the transcription and degradation rate parameters on average expression of the observables (mRNA levels of clock genes) for a complete circadian cycle (24 hrs) after reaching a stable oscillation. A 10% increase in the transcription rate parameter of each gene demonstrated a direct correlation with the corresponding mRNA level. This phenomenon was evident across the entire network, as elevating the transcription rate parameter resulted in a proportional rise in the average expression levels of the associated observables (Fig 14 B). Moreover, the intricate interplay of positive and negative feedback regulations within the gene network was reflected in the sensitivity analysis of transcription and degradation rates. For instance, an increase in the transcription rate parameter of the *Cry* gene (v_{Cs}) led to a decrease in the average expression of *Per* mRNA due to the delayed inhibition imposed by *PER-CRY_n*. Similarly, elevating the transcription rate parameter of *Bmal1* (v_{Bs}) resulted in a decrease in the average expression of *Rev-erb* and an increase in the average expression of *Per* and *Cry*, attributable to the positive feedback loop involving *Bmal1*, *Per*, and *Cry*, alongside the negative feedback loop involving *Bmal1* and *Rev-erb* (Fig 14 A-B).

Conversely, a 10% increase in the degradation rate parameters exhibited an inverse correlation with their corresponding observables, causing a decline in the average expression levels of their associated genes (Fig 14 C). Notably, an increase in the degradation rate parameter of *Bmal1* (d_{mB}) led to an increase in the expression level of the *Rev-erb* gene, owing to the negative

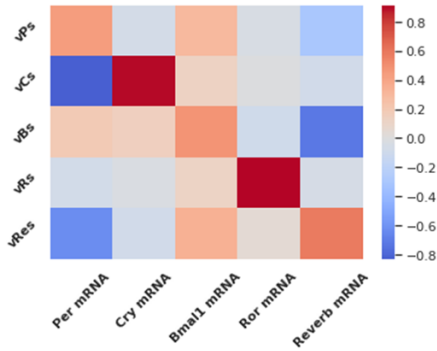
feedback loop involving Bmal1, PER-CRY, and Rev-erb genes. Similar relationships between degradation rate parameters and observables were elucidated across the network (Fig 14 C).

This analysis revealed that the clock gene network demonstrated greater sensitivity to changes in transcription rates compared to degradation rates. Notably, among all genes analyzed, Bmal1 exhibited the highest sensitivity to degradation rates across the entire network, despite being directly regulated solely by Ror and Rev-erb. In summary, our study sheds light on the intricate regulatory dynamics of the circadian clock gene network, highlighting the differential effects of transcription and degradation rate parameters on gene expression dynamics. These findings deepen our understanding of the underlying mechanisms governing circadian rhythm regulation and underscore the complex interplay of feedback mechanisms within the gene network.

A



B



C

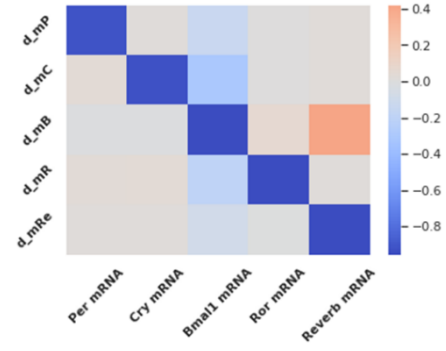


Figure 14: (A) Schematic diagram of the mammalian clock network with its transcription and degradation parameters. (B) Sensitivity analysis of transcription rate parameters. The transcription rate parameter for the observables (mRNA levels of clock genes) were optimized and the dynamics revealed a positive correlation between the transcription rate parameters and the average mRNA level of the respective gene. (C) Sensitivity analysis of degradation rate parameters. Optimization of the degradation rate parameter for the observables (mRNA levels of model clock genes) revealed a negative correlation between the degradation rate parameters and their respective average mRNA. Also, an increase in the degradation rate parameter of the Bmal1 gene, d_mB, led to an increase in the expression level of the Reverb gene due to the negative feedback loop between Bmal1, PER-CRY and Reverb gene.

Increase in the transcription rate parameters lead to a non-monotonic decrease in the oscillatory period while increasing the amplitude of the oscillation.

The circadian clock genes exhibit oscillatory behavior with periods ranging from 23.5 to 24.5 hours¹⁰. While the oscillatory period is consistent across all genes, the amplitude, defined as the absolute difference between the peak and trough levels of the oscillation, varies from gene to gene due to the different biochemical reactions involved in the transcription of each gene. A systematic

exploration of the transcription rate parameters within the range of 0 to 5 a.u revealed a non-monotonic change in the oscillatory period. Specifically, the periods of the circadian clock genes *Per* and *Reverb* initially increased linearly at very low transcription rates, but subsequently decreased monotonically with increasing transcription rate parameters for *Per* (vPs) and *Reverb* (vRes) (Fig 15 A). Conversely, the oscillation period of the *Ror* gene exhibited a monotonic increase with an increase in its corresponding transcription rate parameter, consistent with previous studies attributing this behavior to the positive feedback loop mediated by ROR¹¹⁶. To investigate the effect of transcription rate parameters on the amplitude of circadian clock oscillations, the same range of parameter values (0-5) was explored. The amplitude of the oscillation was monitored across all clock genes in the model. The analysis revealed a consistent trend of monotonic increase in the amplitude of oscillation for all clock genes as their corresponding transcription rate parameters were increased within the specified range. This observation suggests a direct relationship between the rate of gene transcription and the magnitude of oscillatory behavior exhibited by the circadian clock genes. Specifically, as the transcription rate parameters were incrementally increased from 0 to 5, the amplitude of oscillation for each clock gene exhibited a continuous and unidirectional increase, without any deviations or fluctuations in the observed trend (Fig 15 B). This monotonic increase in amplitude indicates that higher transcription rates lead to more pronounced oscillations, with greater deviations from the mean value. It is noteworthy that the rate of increase in amplitude may vary among different clock genes, potentially due to the specific regulatory mechanisms and feedback loops involved in their respective transcriptional processes. However, the overall trend of a monotonic increase in amplitude with increasing transcription rate parameters was consistently observed across all clock genes in the model. These findings highlight the crucial role of transcriptional regulation in modulating the oscillatory

dynamics of the circadian clock network. By adjusting the transcription rate parameters, the amplitude and the period of circadian oscillations can be effectively modulated, potentially influencing downstream processes and physiological rhythms governed by the circadian clock.

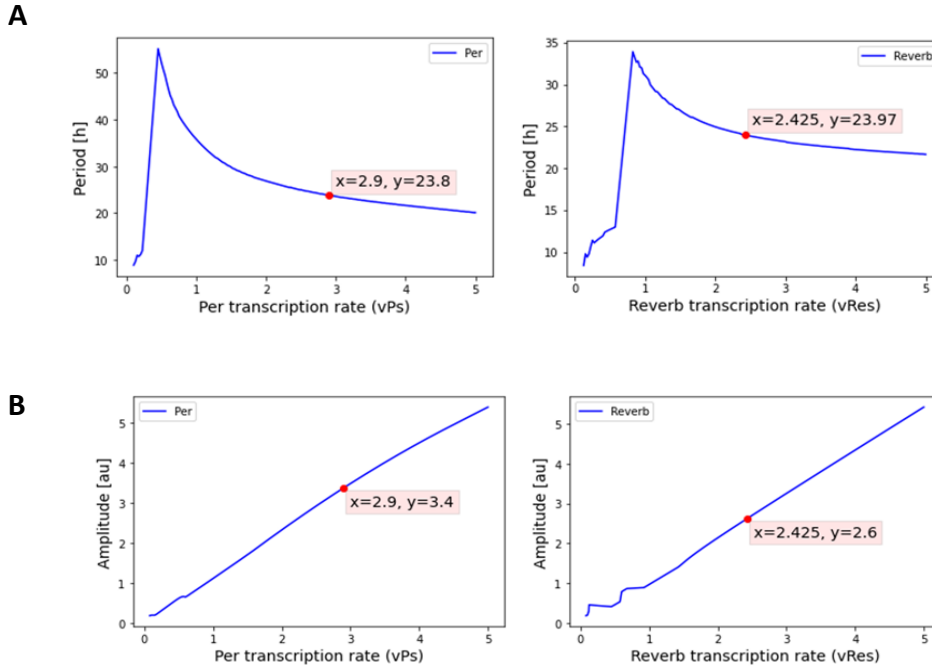


Figure 15: A non-monotonic period and amplitude dynamics for observables Per mRNA and Reverb mRNA as gradient of the transcription rate parameters applied in the model. (A) Oscillation period changes with respect to changes in transcription rate parameter for Per mRNA and Reverb mRNA. The oscillatory period exhibits a monotonic decrease with an increase in the transcription rate parameter. The red dot corresponds to the wild-type parameter value from the simulated model for the transcription rate parameter for Per and Reverb ($x=2.9$ and $x=2.425$) respectively, and their corresponding oscillatory period ($y=23.8$ h and $y=23.97$) (B) Oscillation amplitude changes with respect to changes in transcription rate parameter for Per mRNA and Reverb mRNA. A monotonic increase of oscillatory amplitude is observed with an increase in the transcription rate parameter. The red dot corresponds to the wild-type parameter value for the transcription rate parameter for Per and Reverb ($x=2.9$ and $x=2.425$) respectively, and their corresponding oscillatory amplitude ($y=3.4$ and $y=2.6$).

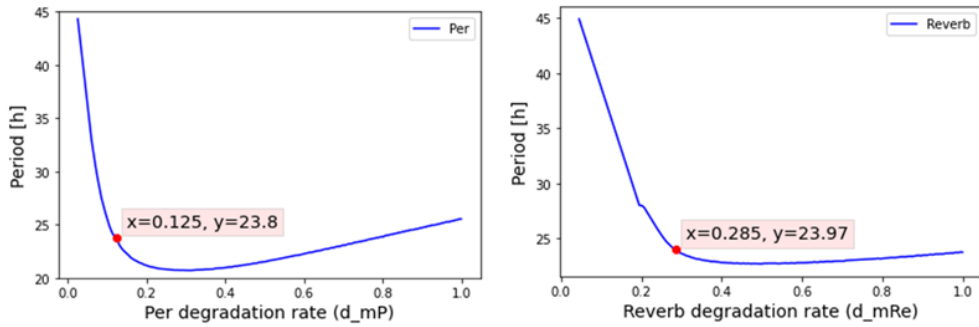
The period shows a concave dependency, along with a monotonic decrease in amplitude as the degradation rate parameter for Per and Reverb is increased.

We conducted a systematic exploration of the parameter space by varying the degradation rate parameters in the model to investigate their impact on the two key characteristics of the circadian

rhythm: the oscillatory period and the amplitude. These parameters govern the rate at which the molecular components of the circadian clock, namely the clock genes and their associated regulatory elements, are degraded or broken down within the cellular environment. Our analysis revealed a non-linear, concave relationship between the degradation rate parameters and the oscillatory period of the circadian clock genes. As the degradation rate parameters were increased incrementally from 0 to 1.0, the period initially decreased, reaching a minimum value, and then increased monotonically with further increments in the degradation rate (Fig. 16 A). This concave pattern was observed for the core clock genes, *Per* and *Reverb*, when their respective degradation rate parameters, d_{mP} and d_{mRe} , were varied independently.

Conversely, our investigation unveiled a monotonic decrease in the amplitude of oscillations across all clock genes in the model, as their corresponding transcription rate parameters were increased (Fig. 16 B). The amplitude, which represents the magnitude of the oscillatory signal, exhibited an inverse relationship with the transcription rate parameters, such that higher transcription rates led to lower amplitudes of oscillation. These findings highlight the intricate interplay between the degradation and transcription rates of the molecular components involved in the circadian clock machinery, and their profound influence on the temporal dynamics and robustness of the circadian rhythm. Notably, the non-linear relationship between the degradation rates and the oscillatory period suggests the existence of an optimal range for these parameters, within which the circadian clock operates most effectively, ensuring the maintenance of a stable and consistent oscillatory pattern.

A



B

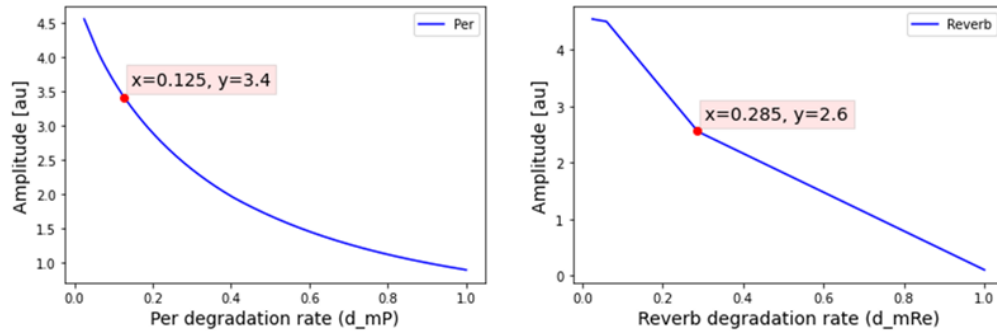


Figure 16: A concave dynamic period for Per and Reverb as degradation rate parameter increases. (A) The oscillation period of both Per and Reverb decreases and then increases giving a concave profile with respect to an increasing degradation rate parameter. (B) Oscillation amplitude for Per and Reverb showed a monotonic decrease with an increase in the degradation rate parameter. The red dots in (A) and (B) denote the wild-type parameter value from the simulated model degradation rate parameter value and the resulting period/amplitude value respectively.

Bifurcation analysis shows that Cry and Ror exhibit Hopf bifurcations with varying transcription rate parameter value.

I further probed the dynamics of the oscillatory rhythm by conducting a bifurcation analysis, which allows for a systematic exploration of how the system's behavior, characterized by properties such as the oscillatory period, amplitude, phase, and average mRNA levels, responds to variations in key parameter values. This analysis was performed using XPPAUT¹¹⁷, a powerful computational

tool for studying dynamical systems. The bifurcation analysis was initially carried out by considering the transcription rates as the primary bifurcation parameters. Specifically, we examined the dependence of the oscillatory dynamics on the transcription rates v_C s for Cry mRNA and v_R s for Ror mRNA. The resulting bifurcation diagrams, which depict the system's behavior as a function of these parameters, revealed a cyclic oscillatory pattern (Fig. 17 A and C). These diagrams illustrate the coexistence of stable and unstable limit cycles, represented by green and blue curves, respectively. The points at which these curves intersect correspond to Hopf bifurcation (HB) points, indicating a transition between steady-state and oscillatory behaviors. By leveraging the information contained within the bifurcation diagrams, I derived the stable and unstable oscillatory periods for both Cry and Ror (Fig. 17 B and D). Notably, the stable and unstable parameter values exhibited a remarkable alignment, both in terms of the oscillatory period and the limit cycle oscillation amplitude, for each gene under consideration. Furthermore, we extended the bifurcation analysis to incorporate the degradation parameters of the observable molecular species, thereby assessing the stability and robustness of the model's rhythmic dynamics under varying degradation rates. The bifurcation analysis not only elucidated the intricate relationships between key parameters and the emergent oscillatory dynamics but also served as a powerful tool for validating the model's capability to capture and maintain robust circadian rhythms. By systematically exploring the parameter space, we gained valuable insights into the regions of stability and instability, as well as the transitions between these regimes, enabling a deeper understanding of the underlying mechanisms governing the circadian clock machinery.

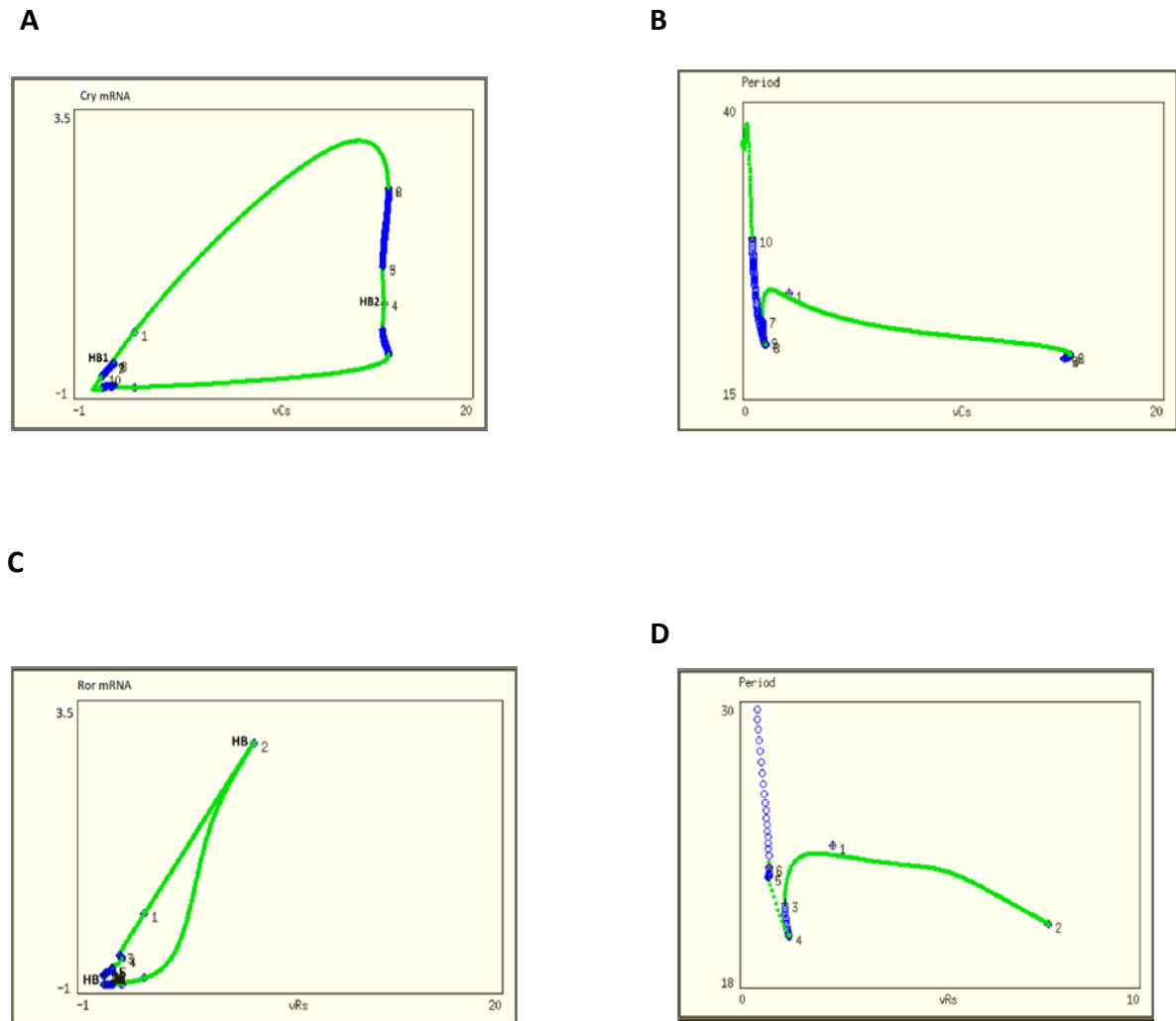


Figure 17: One parameter bifurcation analysis of the model rhythmic dynamics. (A&C) The bifurcation diagram of the transcription rate parameter of Cry mRNA, (vCs) and Ror mRNA, (vRs). The green and blue lines represent stable unstable limit cycles respectively. HB denotes the Hopf bifurcation points in the bifurcation diagram. (B&D) Bifurcation diagram showing the periodic stability of the transcription rate parameter of Cry mRNA, (vCs) and Ror mRNA, (vRs). The green line represents stable period and the blue line represent unstable period of the model.

Model parameter estimation

I used parameter estimation to calibrate the mathematical model of the circadian clock against experimental data obtained from single-nuclei RNA-sequencing of mouse liver cells. The experimental data was acquired from male C57BL/6 mice housed under a 12:12 light:dark cycle, ensuring synchronization with the environmental circadian rhythm¹⁰². Parameter estimation was carried out with the AMIGO2^{113,114} software package, which implements advanced algorithms for optimizing model parameters to achieve the best possible agreement between model predictions and experimental observations. Specifically, we aimed to optimize the transcription and degradation rate parameters of the model to fit the hepatic single-nuclei RNA-sequencing data. The goodness of fit between the model predictions and the experimental data was quantified by a cost function, which measures the discrepancy between the observed values and the values predicted by the model for a given set of parameters. The parameter estimation algorithm seeks to minimize this cost function, effectively identifying the parameter values that yield the closest agreement between the model's output and the empirical observations. A smaller overall value of the cost function indicates a better match between the model's predictions and the available data. I defined the cost function as a maximum (log-) likelihood function, which is a suitable choice given the availability and nature of the measured noise in the experimental data. The initial values for the transcription and degradation rate parameters to be estimated were set to the wild-type values used in the original model simulation. Additionally, we imposed upper and lower bounds on these parameters based on values reported in the literature on circadian system modeling^{37,38,118}. Mathematically, I formulated a non-linear programming problem (NLP) with algebraic constraints, aiming to find the optimal transcription and degradation rate parameters that minimize the cost function. This formulation allows for the incorporation of additional constraints, such as parameter

bounds and other physiological considerations, ensuring that the estimated parameter values are biologically plausible and consistent with prior knowledge. By employing this parameter estimation approach, I was able to refine our mathematical model to better capture the dynamics observed in the experimental data, ultimately enhancing our understanding of the underlying circadian clock machinery and its regulatory mechanisms.

Estimation of transcription and degradation rate parameters.

To assess the predictive accuracy of our mathematical model, I utilized single-nuclei gene expression data obtained from male C57BL/6 mice at specific time points: 2, 4, 8, 12, 18, and 24 hours. My focus was on the core circadian genes incorporated in the model: *Per*, *Cry*, *Bmal1*, *Ror*, and *Reverb*. The normalized RNA-sequencing data replicates for these genes constituted the experimental dataset against which I optimized the transcription and degradation rate parameters. To represent the overall gene expression levels, I averaged the expression values of individual isoforms for each gene. Given the large scale and non-linear nature of our model, local optimization techniques can often converge to suboptimal local minima, failing to identify the globally optimal solution. To overcome this challenge and efficiently locate an accurate global optimum, I employed the enhanced Scatter Search (eSS) hybrid global-local optimization algorithm available in the AMIGO toolbox^{119,120}. This advanced algorithm combines global and local search strategies, enabling a more comprehensive exploration of the parameter space and increasing the likelihood of finding the globally optimal solution. In the optimization process, I targeted five transcription rate parameters and five degradation rate parameters, each associated with one of the observable variables in the model. Notably, the optimized parameter values exhibited a positive correlation with the nominal wild-type parameter values used in the initial model formulation. Consequently, the model predictions generated using the optimized parameter

set demonstrated a positive correlation with the experimental measurements for all five observable variables (Fig. 18 A and B). This strong agreement between the model's output and the empirical data validates the effectiveness of my parameter estimation approach and highlights the model's ability to accurately capture the dynamics of the circadian clock machinery. By leveraging advanced optimization techniques and incorporating experimental data, I successfully calibrated the model, enhancing its predictive capabilities and enabling a more accurate representation of the underlying biological processes governing the circadian rhythm.

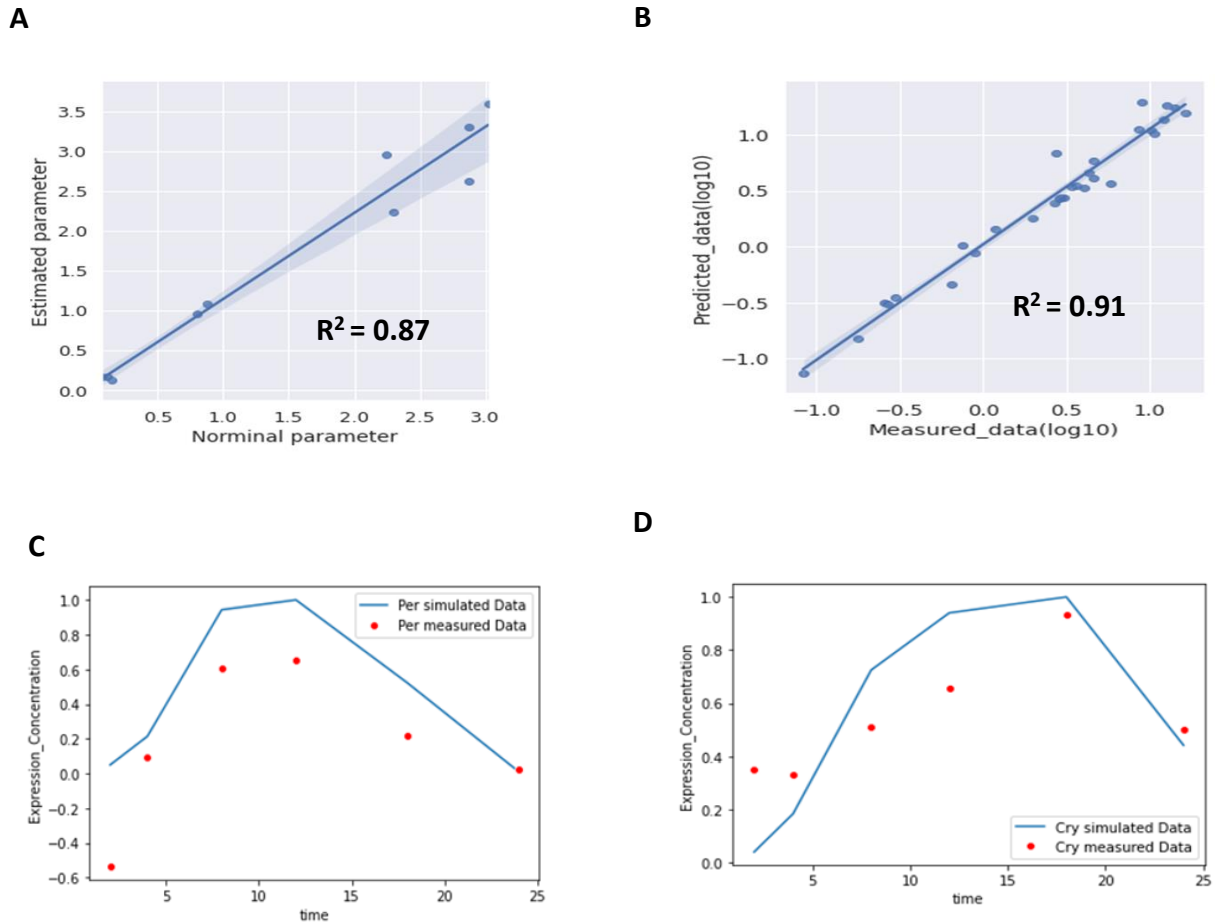
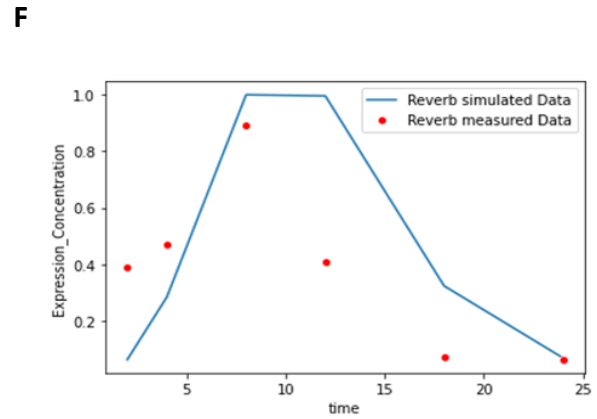
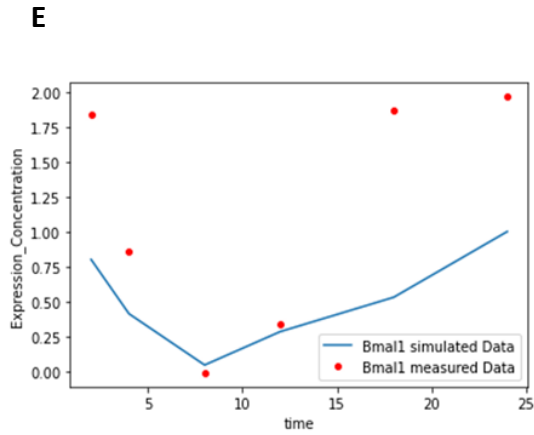


Figure 18: The predicted model results and experimental data fit for observables using maximum (log-) likelihood function. (A) Correlation between predicted transcription and degradation and wild type transcription and degradation rate parameter values. (B) Correlation between model predictions and measured data.

Figure 18 (cont'd)

(C-F) The fitted curves for the model variables (Per, Cry, Bmal1 and Reverb). The red dotted line indicates the experimental data at timepoints 2,4,8,12,18 and 24 h. The solid blue line indicates the predicted results from using maximum (log-) likelihood function to fit the model to the experimental dataset.



CHAPTER 4: SPATIAL-TEMPORAL PERTURBATION OF THE LIVER LOBULE BY 2,3,7,8-TETRACHLORODIBENZO-*P*-DIOXIN

INTRODUCTION

The mammalian liver is a highly organized organ with a unique spatial architecture that facilitates its various metabolic functions. The fundamental structural unit of the liver is the hepatic lobule, primarily composed of hepatocytes, the parenchymal liver cells. Hepatocytes are arranged along the portal-central axis of the lobule, forming a gradient from the portal triad (comprising the portal vein, bile duct, and hepatic artery) to the central vein²⁹. This spatial organization of hepatocytes within the liver lobule is closely linked to the compartmentalization of gene expression and subsequent metabolic activities. Hepatocytes situated closer to the portal triad exhibit higher expression levels of genes involved in gluconeogenesis, oxygen utilization, and β -oxidation processes^{33,121}. In contrast, hepatocytes located nearer to the central vein display increased expression of genes associated with glycolysis, lipogenesis, and xenobiotic metabolism mediated by cytochrome P450 enzymes¹²². This spatial organization of gene expression and metabolic function along the porto-central axis of the hepatic lobule is known as liver zonation. This is a continuous pattern of concentric layers of hepatocytes, reflecting the specialization of different regions of the liver lobule for specific metabolic tasks^{29,33,95}. Liver zonation is established and maintained by a complex interplay of chemical cues, metabolic gradients, and cell-to-cell interactions. The establishment of metabolic gradients is facilitated by the direction of blood flow from the portal triad to the central vein, creating spatial gradients of nutrients and metabolites along the porto-central axis. Additionally, circadian rhythms and molecular signaling from core clock genes contribute to the temporal organization of metabolic processes within hepatocytes. The canonical master regulatory pathway governing liver zonation is the Wnt/ β -catenin signaling pathway. Wnt proteins bind to Frizzled receptors, leading to the phosphorylation of the β -catenin

degradation complex by lipoprotein receptor-related proteins. This event causes the dissociation of β -catenin from the degradation complex, allowing its translocation to the nucleus, where it activates transcriptional regulators that potentiate the zonation process^{34,123,124}.

The temporal regulation of liver functions is intricately governed by the biological phenomenon known as the circadian rhythm. This temporal compartmentalization ensures that metabolic and biological processes within the liver, such as glycolysis and gluconeogenesis, are synchronized with the feeding and fasting cycles of the organism^{28,29,121}. The molecular machinery driving circadian rhythms is a complex network of circadian genes and proteins interconnected through intricate negative and positive feedback loops. The core components of this molecular oscillator include the CLOCK and ARNTL (or NPAS2) proteins, which form heterodimeric complexes. These CLOCK-ARNTL or NPAS2-ARNTL complexes bind to E-box motifs in the promoter regions of downstream circadian genes, such as *Per*, *Cry*, *Ror*, and *Reverb*, activating their transcription. The translated PER and CRY proteins form the PER-CRY complex in the cytoplasm, which subsequently inhibits the binding of the CLOCK-ARNTL or NPAS2-ARNTL complexes, thereby creating a negative feedback loop. Additionally, the REV-ERB protein competes with the ROR protein to inhibit the transcription of the *Arntl* gene, further modulating the circadian rhythm^{3,75}.

A recent study by Droin et al²⁹. demonstrated that the spatial and temporal axes of the hepatic lobule interact with one another. Consequently, many established zoned pathways exhibit rhythmic patterns of gene expression. Among these pathways are gene sets involved in drug metabolism, suggesting a potential interplay between the spatial zonation, temporal rhythmicity, and chemical perturbations within the liver. Notably, while rhythmicity impacts the core pathways that determine zonation, the converse is not true, as the core circadian clock genes themselves do

not exhibit zonation patterns. This observation highlights the complexity of the interplay between the spatial and temporal axes in regulating liver function. The overlap between the temporal and spatial axes, particularly in the context of drug metabolism pathways, suggests the existence of a third dimension to consider when describing liver function: chemical perturbation. This third axis represents the influence of exogenous and endogenous chemical compounds on the intricate spatial and temporal dynamics within the liver.

2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) is a persistent environmental toxicant known for its detrimental effects on various biological functions. Upon entry into the body, TCDD binds to the aryl hydrocarbon receptor (AHR), initiating a cascade of events that lead to toxic effects. Upon binding to TCDD, the AHR undergoes a conformational change, translocates to the nucleus, and binds to specific DNA sequences called dioxin response elements (DREs) located in the proximal promoter regions of target genes, particularly the cytochrome P450 (CYP) genes^{28,51,125}. The binding of the TCDD-activated AHR to DREs modulates the transcription of these target genes, leading to the subsequent inhibition or activation of downstream genes involved in various biological processes. Notably, TCDD exposure has been shown to disrupt both the spatial and temporal components of hepatic function within the liver lobule, ultimately contributing to the development of several liver-related diseases and disorders. Exposure to TCDD has been implicated in the pathogenesis of autoimmune hepatitis, non-alcoholic fatty liver disease (NAFLD), cardiovascular disease, bipolar disorder, obesity, and cancer. The mechanisms underlying these adverse effects involve the disruption of the intricate spatial organization and temporal regulation of gene expression within the liver lobule^{28,89,124}. Specifically, sub-chronic exposure to TCDD has been demonstrated to ablate or significantly dampen the oscillations of most core circadian clock genes, altering their phase and leading to the disorganization of the

rhythmic expression patterns of various zoned gene biomarkers. This disruption of the spatial and temporal axes of hepatic function by TCDD can have far-reaching consequences on liver homeostasis and overall organismal health¹²⁶.

Despite our current understanding of how TCDD-activated AHR may impact these organizational pathways, the specific mechanisms underlying the interaction between these regulatory axes remain elusive. Furthermore, the acute effects of TCDD-activated AHR on the temporal (rhythmicity) and spatial (zonation) components of the hepatic lobule are not well defined. To elucidate the transcriptional impact of acute TCDD exposure on hepatocyte rhythmicity and zonation, we employed hepatic single-nuclei RNA-sequencing data from male C57BL/6 mice administered a single dose of 30 µg/kg TCDD (or sesame oil vehicle) at time-point 0. Liver samples were collected and snap-frozen at various time-points (2, 4, 8, 12, 18, and 24 hours) after TCDD administration. Zonation within the hepatic lobule was inferred from pseudo-space and benchmarked against zonal gene biomarkers from previous studies by Halpern et al⁹⁵. and Droin et al²⁹. Genes were then classified based on the effects of TCDD, zonation, temporal regulation, or a combination of these variables using a mixed-nonlinear effect model. Our analysis revealed that more genes were regulated by zonation independently than by rhythmicity and TCDD alone. Notably, a larger number of genes were regulated by a combination of TCDD and zonation than by a combination of TCDD and rhythmicity. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) analysis of dose-affected categories revealed the interaction of dioxin response elements (DREs), E-box motifs, and AHR in most of the genes affected by TCDD-activated AHR. Overall, our study demonstrated that TCDD exerts a significant effect on hepatic zonation and rhythmicity even at acute exposure levels. The interplay between TCDD, zonation, and rhythmicity in regulating gene expression highlights the complexity of the regulatory

mechanisms governing hepatic function. Importantly, our findings suggest that the disruption of zonation by TCDD may be a more prominent event than the disruption of rhythmicity, as evidenced by the greater number of genes regulated by a combination of TCDD and zonation. This observation underscores the potential impact of environmental toxicants on the spatial organization of hepatic function and the associated consequences for liver homeostasis. Furthermore, the interaction between DREs, E-box motifs, and AHR in the genes affected by TCDD-activated AHR suggests a potential crosstalk between the AHR signaling pathway and the molecular machinery governing circadian rhythms. This crosstalk may contribute to the observed disruption of both zonation and rhythmicity upon TCDD exposure.

METHODS

Single-nuclei RNA-seq expression dataset and preprocessing.

Cholico et al¹²⁷. conducted a time-series experiment to investigate the effects of TCDD on hepatic gene expression in male C57BL/6 mice. The experimental design involved treating the mice with either sesame oil (vehicle control) or a single dose of 30 µg/kg TCDD via gavage. At specific time points (2, 4, 8, 12, 18, and 24 hours) post-treatment, the animals were euthanized by CO₂ asphyxiation, and their livers were immediately collected and snap-frozen for subsequent single-nucleus RNA-sequencing (snRNA-seq) analysis. The single-nuclei RNA-seq dataset generated from this experiment was processed and analyzed using methodologies outlined in previous studies^{124,126}. Specifically, the clustering and cell type annotation of the single-nucleus transcriptomes were performed to identify and characterize the various hepatic cell populations present in the samples.

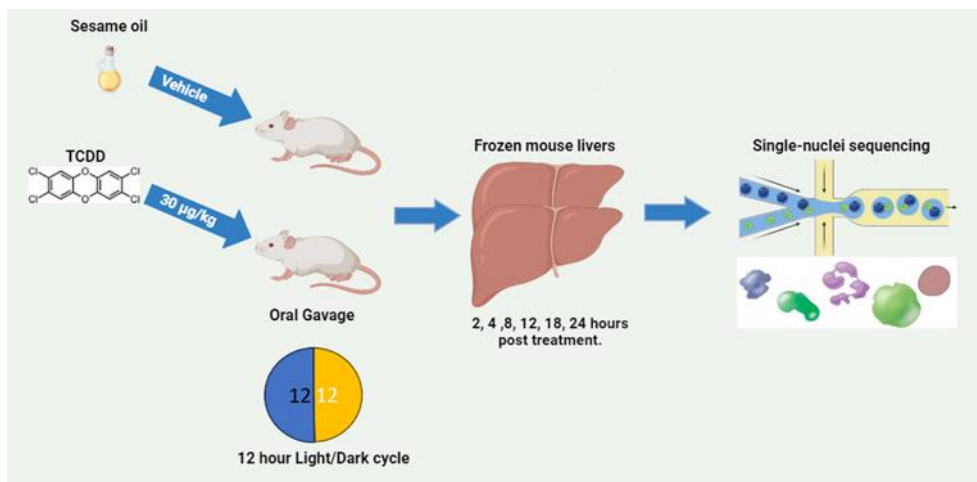


Figure 19: Extraction and sequencing of single nuclei data from male C57BL/6 mice. Male C57BL/6 mice housed in a room with a 12:12 light dark cycle was gavaged with single dose of sesame oil (as vehicle) or 30 µg/kg of TCDD (as treated). Their livers were harvested 2, 4, 8, 12, 18, 24, hours post treatment and snap frozen. The snap frozen livers were sequenced to acquire the single nuclei sequencing dataset.

We employed a comprehensive preprocessing pipeline to prepare the single-nucleus RNA-sequencing (snRNA-seq) dataset for downstream analysis. The preprocessing of the dataset was conducted using the scanpy package in Python¹²⁸, a bioinformatics tool for analyzing single-cell RNA sequencing (scRNA-seq) data. Raw counts from the dataset were first normalized to the median total cell count using the `normalize_total` function, ensuring consistency across samples. Subsequently, a log transformation with a pseudocount was applied using the `log1p` function to mitigate the effects of extreme values and bring the data into a more interpretable range for subsequent analyses. Next, the preprocessing pipeline focused on filtering cells and genes to remove low-quality or uninformative data points. Since the study aimed to investigate the effects of TCDD on hepatocytes, the first filtering step involved removing all non-parenchymal cell types, retaining only hepatocytes in the dataset. This step ensures that the analysis focuses specifically on the cell type of interest. Subsequently, low-quality hepatocytes were filtered out based on two criteria: 1) hepatocytes with fewer than 1,500 counts, and 2) hepatocytes with fewer than 200

detected genes. These thresholds were chosen to exclude cells with low sequencing coverage or poor transcriptome quality, as they may introduce noise and bias into the analysis. In addition to cell filtering, gene filtering was also performed. Genes that were expressed in fewer than 200 cells were removed from the dataset. This step helps to reduce the computational burden and focus the analysis on genes with sufficient expression levels across the cell population. We adopted a non-standard preprocessing step due to the lower formation of clusters observed during batch correction analysis. This step highlights the importance of tailoring the preprocessing pipeline to the specific characteristics of the dataset and the research questions being addressed. Finally, the identification of highly variable genes (HVGs) was accomplished using the `highly_variable_genes` function from the scanpy package. HVGs are genes that exhibit substantial variability in expression levels across cells, potentially reflecting biologically relevant differences. These genes are typically used as input for downstream analyses, such as dimensionality reduction and clustering, as they capture the most informative aspects of the transcriptomic landscape.

Batch correction using scVI.

To mitigate sample-specific batch effects present in the snRNA-seq data, we employed scVI¹²⁹ a variational autoencoder model designed for single-cell RNA-sequencing data analysis. Variational autoencoders are a type of deep learning model that can learn complex, nonlinear representations of high-dimensional data, making them well-suited for batch correction and other tasks in single-cell genomics. To account for batch effects arising from different biological samples, we assigned a unique batch label to the cells originating from each individual sample. These sample-specific batch labels were then provided as input to the scVI model during training. By incorporating batch information during the model's training process, scVI learns to disentangle the biological signal from the technical artifacts introduced by batch effects. The scVI model was trained using the

hyperparameters and architecture configurations specified in Tables 2.1 and 2.2, which detailed the model's neural network architecture, regularization techniques, optimizer settings, and other training specifications. These hyperparameter settings were deemed suitable for the given snRNA-seq data and batch correction task.

<i>Hyperparameter</i>	<i>Value</i>
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 3.1: Hyperparameters for scVI’s variational autoencoder model.

<i>Hyperparameter</i>	<i>Value</i>
Training epochs	46
Learning rate	0.001
Learning rate decay	10^{-6}
Optimizer	Adam
Optimizer epsilon	0.01

Table 3.2: Hyperparamters for scVI’s variatonal autoencoder training.

Layer calculations.

We utilized the latent space representation of the normalized cell counts obtained from the single-cell variational inference (scVI) method as input into the diffusion maps algorithm. Diffusion maps were generated using the diffmap function from the scanpy Python package. This technique is a non-linear dimensionality reduction method that captures the underlying geometric structure of the data in a low-dimensional representation. The second component of the diffusion maps representation was extracted and min-max scaled to generate a pseudo-space metric. This metric was oriented such that genes with central zone enrichment had the highest values, while genes with portal zone enrichment had the lowest values. The rationale behind this approach is to capture the transcriptional heterogeneity along the central-portal axis of the liver lobule. The cells were then divided into five equal bins based on their pseudo-space metric values, representing transcriptional zones along the central-portal axis. Each bin contains cells within one-fifth of the full pseudo-space range (i.e., cells in bin i have pseudo-space values between $\frac{i-1}{5}$ and $\frac{i}{5}$). This binning approach allows for the identification of genes with varying expression patterns across the central-portal axis. For each treatment condition, time point, and bin (layer), the raw counts were

summed across the constituent cells. This step aggregates the gene expression data across cells within each bin, effectively capturing the transcriptional profile of each zone. These count sums then underwent normalization using the `computeSumFactors` function from the `scran` R package to obtain "Normalized Counts." This normalization step is crucial to account for technical biases and variations in sequencing depth, ensuring that the gene expression values are comparable across different samples and conditions. Genes were then categorized by zone based on the bin (layer) showing their maximum Normalized Counts. Genes with peak expression in bins 1 and 2 were classified as central zone genes, as they exhibit the highest expression levels in regions closer to the central vein. Bin 3 genes with mid-range pseudo-space values were termed mid-lobular, representing genes with intermediate expression patterns. Finally, portal zone genes displayed maximal Normalized Counts in bins 4 and 5, indicating their enrichment in regions closer to the portal vein.

Design of Non-Linear Mixed Effect Model.

We devised an analytical framework by constructing a non-linear mixed effect model (NLMEM) using the `MixedLM` class in the `statmodels` Python package to investigate the effects of various factors on gene expression. Mixed effect models are statistical models that incorporate both fixed and random effects, allowing for the analysis of data with hierarchical or nested structures. In this study, the NLMEM models were meticulously designed to incorporate various factors, enabling the examination of the impact of TCDD exposure (D), rhythmicity (R), and zonation (Z) on gene expression dynamics. Table 2.3 provides descriptions of the individual terms representing each factor in the NLMEMs. These terms capture the specific aspects of each factor that influence gene expression. For instance, the term representing TCDD exposure may include parameters related to the dose, duration, or time since exposure, while the term for rhythmicity may incorporate

parameters related to the amplitude, period, and phase of the oscillations. The specific equations for the NLMEM classes outlined in Table 2.3 are presented in Table 2.4. These equations mathematically represent the relationships between the factors (D, R, and Z) and gene expression, incorporating both fixed and random effects. Fixed effects are parameters that are constant across all observations, while random effects account for the variability among different groups or levels within the data. The NLMEM approach offers several advantages over traditional linear models. First, it allows for the incorporation of non-linear relationships between the predictors and the response variable, which is particularly relevant for biological systems where responses may be non-linear or exhibit complex patterns. Second, the mixed effect structure accounts for the hierarchical or nested nature of the data, where observations may be correlated within groups or clusters. This is crucial when analyzing gene expression data, where measurements from the same individual or experimental unit may be correlated.

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

Table 3.3: Terms for mixed linear effects models. Each term is denoted by the name and its effect. D is TCDD (Dioxin) Influence, R is rhythmicity, and Z is zonation. t is the time in hours after

Table 3.3 (cont'd)

treatment. l is the layer of the liver lobule. ω is the conversion factor between t and radians which is equal to $\frac{2\pi}{24}$.

<i>Class</i>	<i>Equation for Model</i>
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)	$y = \beta_0 + \beta_1 D * (\beta_2 R_{Sin} + \beta_3 R_{Cos} + \beta_4 Z_{P1} + \beta_5 Z_{P2})$
DxZxR	$y = \beta_0 + \beta_1 D * (\beta_2 R_{Sin} + \beta_3 R_{Cos}) * (\beta_4 Z_{P1} + \beta_5 Z_{P2})$

Table 3.4: Equations for each non-linear mixed effects model used for classification.

Implementation of the NLMEM was almost identical to Droin C et al²⁹. These equations were fit to normalized count of each individual gene using the Nelder-Mead optimization algorithm¹³⁰. A noise offset ($\sigma_0 = 0.15$) was added to the data to avoid overfitting. Equations with the smallest overall Bayesian information criterion¹³¹ (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 was when

models tied with one another. Like for this was in ties, which we defined 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:

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

Differential Rhythmicity and Zonation.

To conduct a comprehensive analysis of differential rhythmicity and zonation between treated and control conditions, gene expression data were fitted to models similar to the approach described in section 2.4. and the Differential RhythmicitY analysis in R (DryR)¹³² framework model in R. DryR (Differential RhythmicitY analysis in R) is a statistical framework based on model selection that is designed to detect and estimate changes in rhythmic parameters (amplitude and phase) and mean expression levels across multiple conditions. This framework leverages the power of rhythmic regression models and likelihood-based model comparison techniques to rigorously analyze time-course gene expression data. The core principle of DryR is to fit gene expression data to a set of competing models that represent different rhythmic behaviors and subsequently compare the goodness-of-fit of these models using likelihood ratio tests. By fitting these models to the time-course data and comparing their respective likelihoods, DryR can identify genes that exhibit rhythmic expression patterns, as well as those that display significant changes in rhythmicity or mean expression levels between conditions.

With respect to the models similar to the approach described in section 2.4, instead of using the Nelder-Mead optimization algorithm, we employed ordinary least squares (OLS) regression for

model fitting by leveraging the statsmodels Python library. The OLS regression method is a widely used technique for estimating the unknown parameters in a linear regression model. By minimizing the sum of squared residuals between the observed data and the predicted values from the model, OLS provides an unbiased and efficient estimate of the model parameters under certain assumptions. Specifically, the `ols` function from `statsmodels.api` was utilized to carry out OLS on each gene's expression values, which were separated into treated and control groups. This separation enabled the quantification of differences in rhythmicity parameters and spatial patterning between the two conditions across various zone regions, supporting subsequent analysis of spatial and temporal expression patterns under treatment conditions. These separated treated and control gene expression values were then fitted to either the R class model (capturing rhythmicity) or the F class model (non-rhythmic flatline) for rhythmicity analysis. The R class model incorporates parameters that describe the oscillatory behavior of gene expression, such as amplitude, period, and phase, while the F class model represents a constant, non-rhythmic expression pattern. Similarly, for spatial zonation analysis, the data were fitted to the Z class model (spatial patterning) or the F class model. The Z class model accounts for the spatial organization of gene expression along the central-portal axis of the liver lobule, capturing the varying expression levels across different zones. To determine if the R class or Z class models, which have more parameters, provided a statistically significantly better fit than the simpler F class model, likelihood ratio tests (also known as χ^2 tests) were employed. The likelihood ratio test is a statistical method used to compare the goodness of fit between nested models, where one model (the null model, in this case, the F class model) is a special case of the other model (the alternative model, either the R class or Z class model). The likelihood ratio test compared the maximum likelihood estimates of the more complex model (R or Z class) to the simpler model (F class),

quantifying if the additional parameters justify the increase in model complexity. This enabled a rigorous identification of differences in rhythmicity and spatial patterning between treated and control conditions for each gene tested. Genes with expression data that were significantly better explained by the R class model compared to the F class model, based on the likelihood ratio test, were classified as exhibiting differential rhythmicity. Similarly, genes better fit by the Z class model were designated as displaying differential zonation. The likelihood ratio test was implemented using the `chi2` function from the `scipy.stats` Python package for this model selection approach. This function calculates the p-value associated with the likelihood ratio test statistic, allowing for the assessment of statistical significance and the selection of the more appropriate model for each gene.

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

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

And we can define the phase as:

$$phase = \omega \arctan(b, a)$$

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

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

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

Statistical Tests.

Differential expression analysis between treatments and comparison of parameter values were performed using the non-parametric Mann-Whitney U test, implemented via the `mannwhitneyu` function in `scipy.stats`. The Mann-Whitney U test is a non-parametric statistical test that is used to compare the distributions of two independent samples. It is particularly useful when the data violates the assumptions of normality or homogeneity of variance, which are required for parametric tests such as the t-test. To determine if there were differences between the distribution of gene zone assignments between gene lists, the Kolmogorov-Smirnov two-sample test was utilized through the `ks_2samp` function. The Kolmogorov-Smirnov test is a non-parametric test that quantifies the distance between the empirical distribution functions of two samples. This test is useful for comparing the overall shapes and distributions of two datasets, rather than just their central tendencies or means. Correlation coefficients and associated p-values were also computed using `scipy.stats` functions. Correlation analysis is a statistical technique that measures the strength and direction of the linear relationship between two variables. The correlation coefficients quantify the degree of association, while the p-values indicate the statistical significance of the observed correlation. All statistical tests leveraged Python-based `scipy.stats` functions, which are part of the SciPy library. SciPy is a widely used open-source Python library for scientific and technical computing, providing a comprehensive collection of mathematical algorithms and functions. For gene set enrichment analysis, the Enrichr API within the `gseapy` package was used to identify enriched pathways and processes. Gene set enrichment analysis (GSEA) is a computational method that determines whether predefined sets of genes (e.g., pathways or biological processes) are overrepresented or underrepresented in a given gene list, compared to a background set of genes. The Enrichr API provides access to a comprehensive collection of gene set libraries,

enabling the identification of enriched biological annotations associated with the genes of interest. Significant pathways were called at a false discovery rate (FDR) threshold of < 0.2 . The FDR is a statistical method used to correct for multiple hypothesis testing, which is a common issue when performing numerous statistical tests simultaneously. By controlling the FDR, researchers can balance the trade-off between false positive and false negative rates, ensuring that the identified enriched pathways are statistically robust. This integrative analysis enabled connecting the observed differential expression and rhythmicity changes to impacted biological functions and processes. By combining the results of differential expression analysis, rhythmicity parameter comparisons, and gene set enrichment analysis, researchers can gain a comprehensive understanding of the transcriptional responses to experimental treatments or conditions. This approach allows for the identification of dysregulated genes, alterations in rhythmic patterns, and the associated biological pathways and processes that may be affected by these changes.

RESULTS

Treatment with 2,3,7,8 Tetrachlorodibenzo-p-dioxin (TCDD) induces an oscillatory pattern in circadian clock genes and elevates the expression of TCDD response genes in acute hepatocyte response.

To investigate the transcriptional impact of acute 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) exposure on the disruption of rhythmicity and zonation in hepatocytes, we utilized single-nucleus RNA-sequencing (snRNA-seq) data from the livers of male C57BL/6 mice, as generated by Cholico et al¹²⁷. The mice were housed in a controlled environment with a 12-hour light/12-hour dark cycle. At the start of the light phase (6:00 AM), the mice received a single oral dose of 30 $\mu\text{g/kg}$ TCDD or a vehicle control (sesame oil) via gavage. Subsequently, the livers were harvested and snap-frozen at specific time points: 2, 4, 8, 12, 18, and 24-hours post-treatment. Following sequencing, the snRNA-seq data was processed, and hepatocytes were identified based on

established hepatic biomarkers from previous studies^{124,126}. After excluding non-hepatocyte cell types and filtering out low-quality hepatocytes with low read counts, a total of 129,373 hepatocytes remained in the dataset. The top 15,000 highly variable genes (HVGs) were selected for downstream analysis. Dimensionality reduction techniques (Batch correction) were applied to the TCDD response genes for data visualization and analysis, as detailed in the methods section (2.2). Uniform Manifold Approximation and Projection (UMAP) analysis was conducted on the hepatocyte population to examine the cell clustering patterns. The UMAP analysis revealed a distinct circular pattern in the hepatocyte clusters, which corresponded to different time points and treatment conditions (Fig. 20 A). This circular pattern suggests an inherent rhythmicity in hepatocyte gene expression, potentially reflecting the circadian rhythm or other cyclical processes within these cells. This circular pattern in the UMAP analysis, indicative of rhythmic gene expression, was not observed in other cell types present in the dataset. Notably, the hepatocyte cluster corresponding to the TCDD treatment group exhibited a distinct separation from the control group, suggesting a significant alteration in the gene expression profiles due to TCDD exposure (Fig. 20 B). To examine the immediate impact of TCDD over time and space, the expression patterns of known circadian clock genes (*Per1*, *Arntl*, and *Cry*) and TCDD response genes (*Cyp1a1* and *Ahrr*) were analyzed. For the circadian clock gene *Per1*, the circular pattern (oscillatory pattern with a peak at the 8-hour time point) was observed (Fig. 20 C), indicating the preservation of rhythmic expression despite TCDD treatment. While acute TCDD treatment impacted hepatic rhythmicity, it did not completely abolish circadian oscillation, as seen in previous studies with sub-chronic TCDD exposure²⁸ (Fig. 2E). This finding suggests that the disruption of circadian rhythms by TCDD may be dependent on the duration and dosage of exposure. As expected, a significant increase in the expression profiles of TCDD response genes (*Cyp1a1* and *Ahrr*) was

observed after treatment (Fig. 20 D and 20 F). Interestingly, the expression of *Ahrr* and *Cyp1a1* achieved saturation at 12 hours post-treatment (Fig. 20 F), while *Cyp1a2* and *Tiparp* reached saturation at or before the 2-hour time point (data not shown). This differential temporal response of TCDD-inducible genes suggests a complex regulatory mechanism underlying the transcriptional response to TCDD exposure. The observed circular pattern in the UMAP analysis, coupled with the analysis of circadian clock genes and TCDD response genes, provides insights into the impact of acute TCDD exposure on hepatocyte rhythmicity and gene expression dynamics. While TCDD treatment altered hepatic rhythmicity, it did not completely abolish circadian oscillations, at least in the acute exposure scenario. Additionally, the varying temporal responses of TCDD-inducible genes highlight the intricate transcriptional regulation mechanisms at play in the liver's response to dioxin exposure.

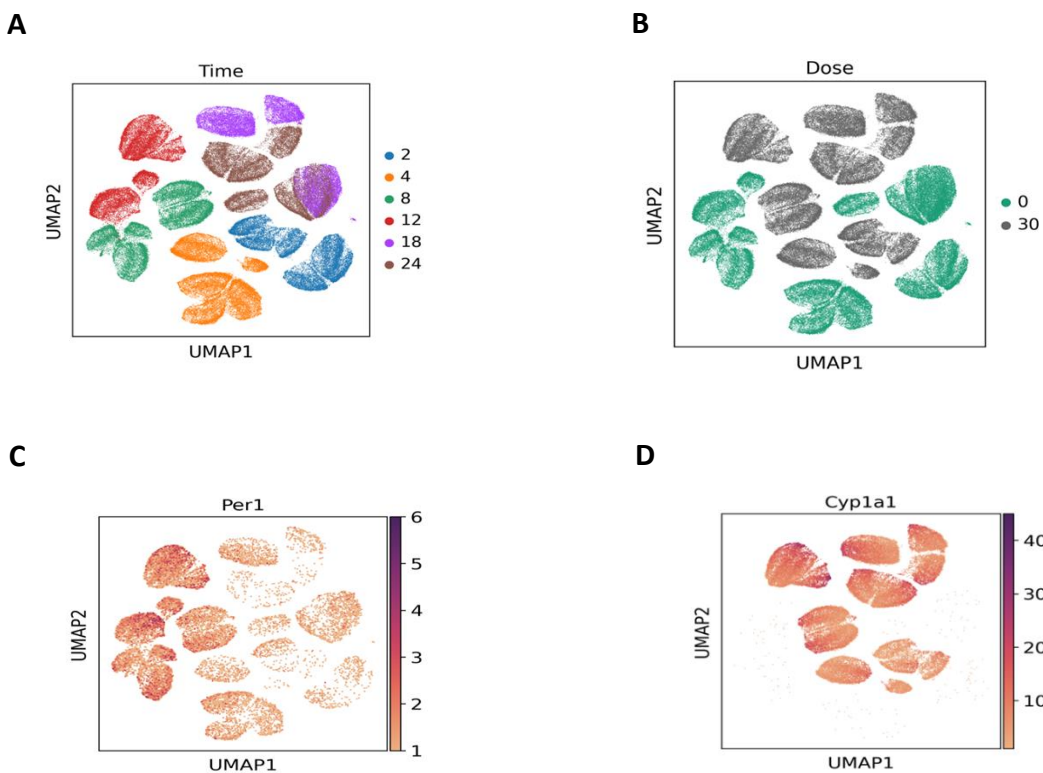
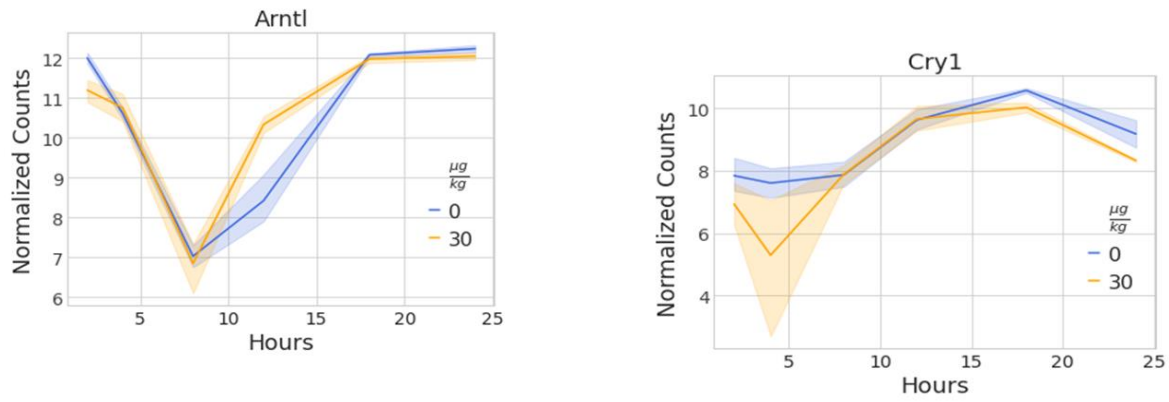


Figure 20: Visualization of acute TCDD in mouse hepatocytes. UMAP of hepatocytes colored by

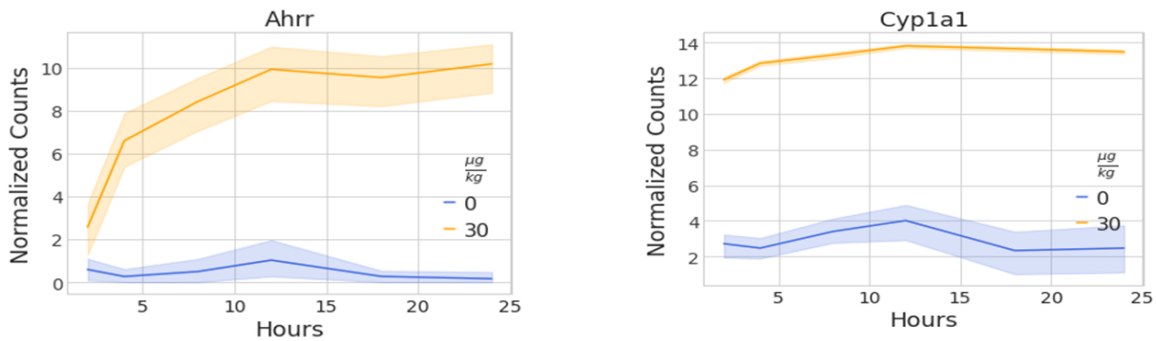
Figure 20 (cont'd)

(A) Time in hours after treatment, (B) Dose in $\mu\text{g/kg}$ of TCDD, (C) Circadian clock gene *Per1* showing peak expression between timepoint 8 to 12 hrs and (D) TCDD activated gene *Cyp1a1* showing almost zero expression at dose 0 $\mu\text{g/kg}$ (vehicle/sesame gavage) and high expression at dose 30 $\mu\text{g/kg}$ (TCDD gavage). Each dot represents a cell. Time series expression of Circadian and TCDD response genes. Normalized counts time series expression of known (E) Circadian clock genes *Arntl* (*Bmal1*), *Cry1* and (F) TCDD activated genes *Ahrr* and *Cyp1a1*. Expression of genes are plotted as a function of sesame oil vehicle (0 $\mu\text{g/kg}$) in blue and TCDD (30 $\mu\text{g/kg}$) in orange.

E



F



Zonation of hepatocyte from single nuclei gene expression profile.

Identifying zonated genes in the liver lobule is a challenging task when using single-nucleus RNA sequencing (snRNA-seq) data, as zonation is not directly measured in these experiments. Instead,

zonation patterns need to be inferred from the gene expression profiles of individual hepatocytes. However, this inference can be confounded by various experimental factors that can distort the zonation profiles. For example, the centrally zonated gene *Cyp1a2*¹³³ is known to be activated by TCDD treatment, while another centrally zonated gene, *Slc1a2*, exhibits a rhythmic expression pattern over time. To accurately infer zonation from hepatocyte gene expression, these confounding factors need to be corrected. To address this challenge, we utilized the approach developed by Nault et al¹²⁴. The first step involved batch correction of the single-cell data to remove variance introduced by TCDD exposure and temporal effects. This step is crucial as it eliminates the influence of these confounding factors on the gene expression profiles, allowing for a more accurate inference of zonation. After batch correction, trajectory inference was performed to calculate the latent zonation value for each individual hepatocyte. This approach leverages the inherent structure and relationships within the gene expression data to infer the underlying zonation patterns. By applying batch correction and trajectory inference, we aimed to disentangle the effects of TCDD treatment and temporal variations from the gene expression profiles, thereby enabling a more accurate inference of hepatocyte zonation. This approach is particularly important in the context of studying the impact of TCDD on zonation, as it allows for the identification of zonated genes that may be affected by dioxin exposure, while accounting for potential confounding factors. To perform batch correction on the single-nucleus RNA sequencing (snRNA-seq) data, we utilized scVI¹²⁹, a variational autoencoder (VAE) method. The choice of a VAE approach was driven by the large size of the dataset, which comprised approximately 130,000 cells after preprocessing. Most single-cell integration tools are not designed to efficiently handle datasets of such a magnitude. However, scVI¹²⁹ was specifically developed to integrate large gene expression atlases across multiple laboratories and to batch correct technical variations, making it well-suited

for this analysis. The batch correction process using scVI aimed to disentangle the technical variation introduced by TCDD treatment and time of harvest from the biological variation in gene expression. By removing these confounding technical factors, the resulting batch-corrected data would provide a more accurate representation of the underlying biological processes, such as hepatocyte zonation and rhythmic gene expression patterns. Visualizations of the batch-corrected data (Fig. 21 A-C) demonstrate the effectiveness of the scVI approach in mitigating the effects of TCDD treatment and time of harvest on the gene expression profiles. This batch-corrected data served as the foundation for subsequent trajectory inference analyses, enabling a more reliable inference of hepatocyte zonation patterns while accounting for the potential confounding factors introduced by the experimental design.

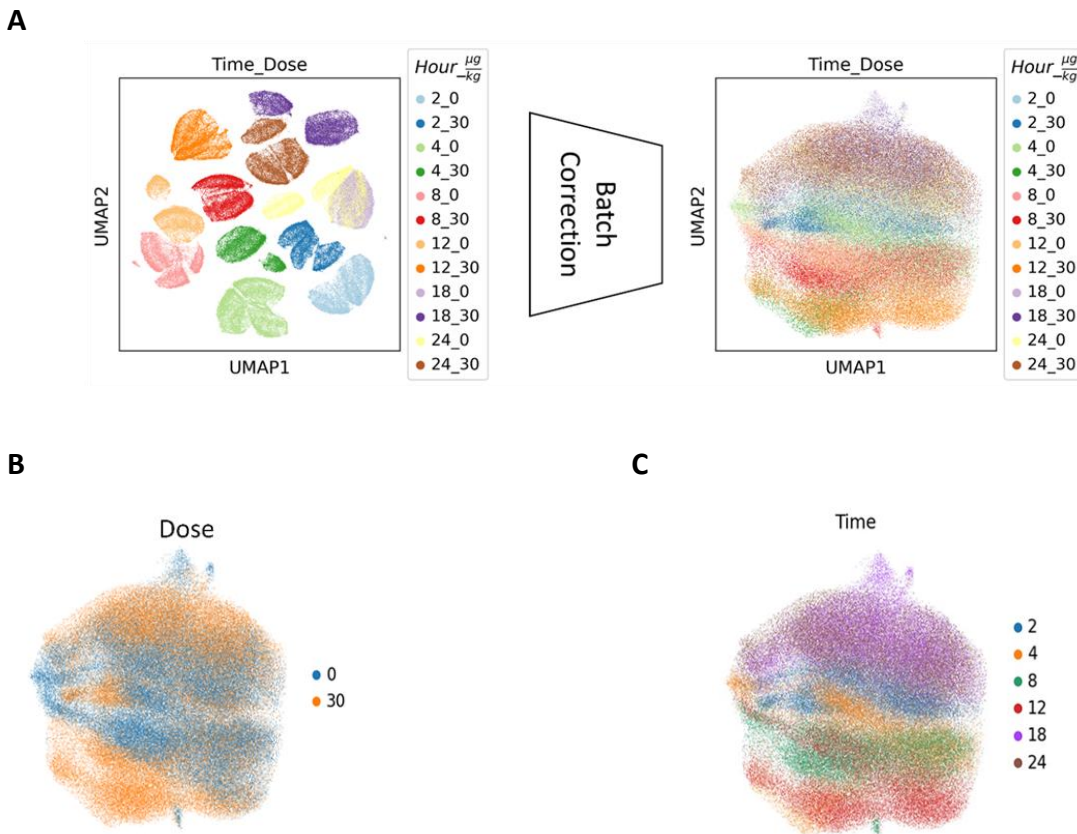


Figure 21: UMAP of batch correction by scVI of single-nuclei RNA sequencing data.

Figure 21 (cont'd)

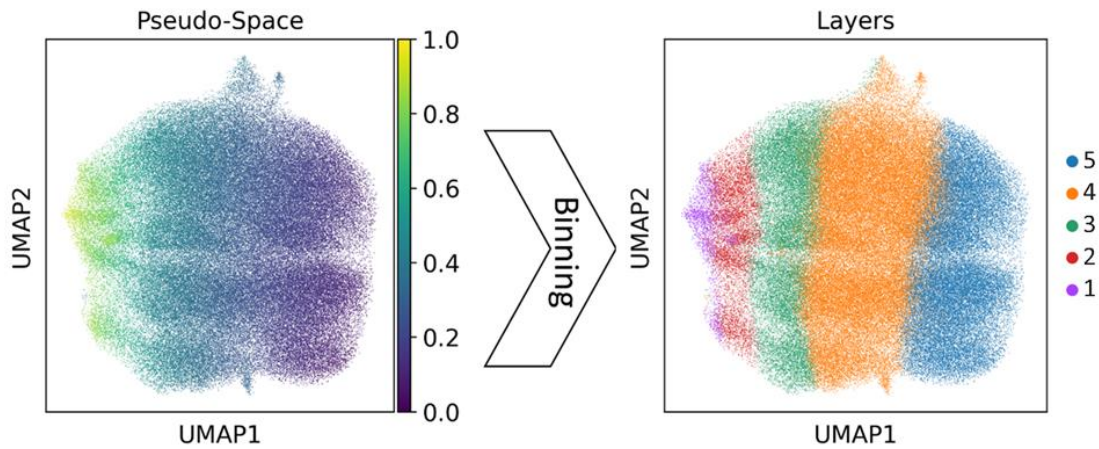
(A) UMAPs of single-nuclei RNA sequencing data and the latent space of scVI (batch correction). Cells are colored by the combined label of hours and TCDD dose treatment. (B) Batch corrected single-nuclei RNA sequencing colored by 0 $\mu\text{g/kg}$ dose treatment (blue) and 30 $\mu\text{g/kg}$ dose treatment (orange). (C) Batch corrected single-nuclei sequencing colored by six (6) time-points in hours after treatment. Each dot represents a cell.

After batch correcting the single-nucleus RNA sequencing (snRNA-seq) data using scVI, we employed a trajectory inference algorithm, diffusion pseudo-time^{134,135} on the latent space generated by scVI to infer the trajectory of the portal-central axis. This approach aimed to capture the zonation patterns within the liver lobule. Utilizing a second component of the diffusion pseudo-time plot (analogous to components of PCA) as our trajectory, we observed most zonal genes follow along the component of either high expression in the portal axis with low expression in the central axis or high expression in the central axis with low expression in the portal axis. Expression values were normalized and reoriented from portal to central expression. We referred to the trajectory inference values as "Pseudo-space," which we defined as an ordering of cells based on how closely they approximate the expression patterns of centrally zoned hepatocytes (represented by a value of 1) and portal zoned hepatocytes (represented by a value of 0). To validate whether the inferred pseudo-space accurately captures liver lobule zonation, we examined the expression profiles of known zoned genes, *Cyp2f2* and *Slc1a2*, across the pseudo-space axis. Significant correlations were observed between the expression levels of these genes and the pseudo-space values of the cells. Specifically, we observed high expression of *Slc1a2* at the central axis (Fig. 22 B) and high expression of *Cyp2f2* at the portal axis (Fig. 22 C), consistent with their known zonation patterns. As a negative control, we analyzed non-zoned genes, such as *Arntl* and *Clock*, which are not expected to exhibit zone-dependent expression patterns¹. As anticipated, these genes did not show significant correlations with the pseudo-space values, further validating the accuracy of the inferred zonation trajectory. Our analysis confirmed that no significant

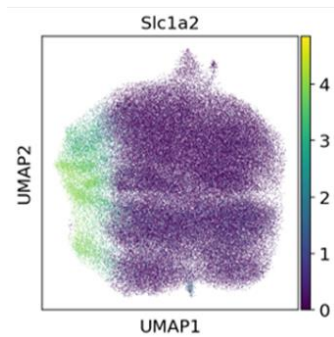
correlations exist between the expression of the non-zonated genes *Arntl* and *Clock* and the computed pseudo-space metric. The expression of these genes was observed to be non-zonated in nature (Fig. 22 D). The presence of strong correlations for known zonated genes, such as *Cyp2f2* and *Slc1a2*, and the absence of such correlations for non-zonated genes, confirms that the computed pseudo-space metric accurately reflects the zonation pattern across the liver lobule. To streamline subsequent analyses, we binned the continuous pseudo-space values into distinct layers (Fig. 22 A). The number of layers used depends on the resolution of the data. For example, Halpern et al⁹⁵. defined fifteen zones, whereas Droin et al²⁹. used only eight zones. In our study, we elected to bin the data into five layers, ensuring at least two thousand cells in each layer. The pseudo-space axis was divided into five equal-length bins. As expected, fewer cells were binned into the first (central) layer compared to the last (portal) layer (Fig. 4A). This observation aligns with the anatomy of the liver lobule, where there is a single central vein and numerous peripheral portal triads, resulting in fewer cells located proximal to the central vein. Using the binned pseudo-space values, we generated pseudo-bulk expression profiles for each treatment, time point, and layer by summing the counts across cells within each combination of these factors. These count sums were then normalized to account for differences in the number of cells comprising each pseudo-bulk profile using size factor estimation¹³⁶. The resulting normalized expression values for each treatment $\times$ time $\times$ layer combination were utilized in downstream analyses and gene classification. The binning of the continuous pseudo-space values into distinct layers facilitated the analysis of zonated gene expression patterns by providing a discrete representation of the spatial organization within the liver lobule. By generating pseudo-bulk expression profiles for each treatment, time point, and layer combination, we could effectively capture the dynamic changes in gene expression across different experimental conditions and spatial locations within the liver lobule. The pseudo-

space values obtained through trajectory inference provide a quantitative measure of hepatocyte zonation, allowing for the identification and analysis of genes exhibiting zoned expression patterns within the liver lobule. By combining batch correction techniques (scVI) and trajectory inference (diffusion pseudo-time), we were able to disentangle the confounding effects of experimental factors and accurately capture the underlying zonation patterns, enabling a more comprehensive understanding of the spatial organization and functional specialization of hepatocytes in response to TCDD exposure.

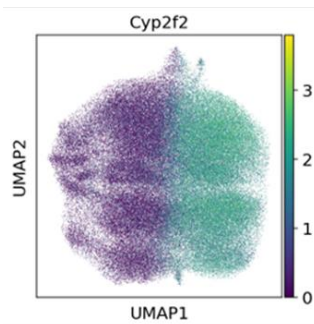
A



B



C



D

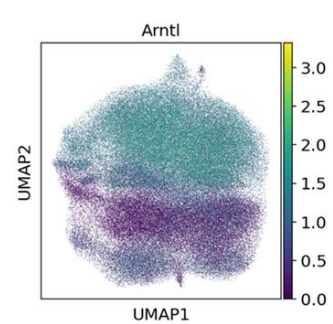


Figure 22: UMAP visualization of pseudo-space binning and zonation. (A) UMAP plot of the pseudo space of hepatocytes binned into zonation layers. The pseudo-space value of 1 represents

Figure 22 (cont'd)

hepatocytes with high pericentral gene expression and 0 represents hepatocytes with high periportal gene expression. Zonation layers are deduced between number 1 (pericentral) and 5 (periportal). UMAP plot of (B) Pericentral gene marker *Slc1a2*, (C) Periportal gene marker *Cyp2f2* and (D) Circadian clock gene *Arntl*. Each dot represents a cell.

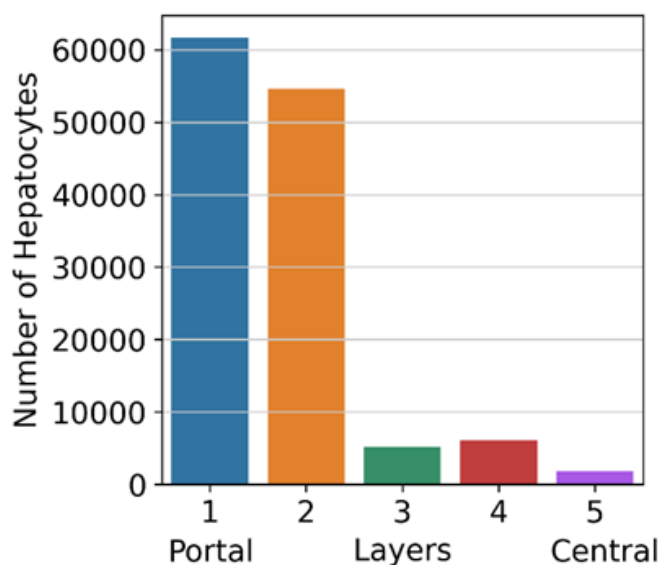


Figure 23: Bar plot of the number of hepatocytes in each inferred layer. Layers with smaller values represent more pericentral hepatocytes and layers with larger values represent more periportal hepatocytes.

Classifying hepatocytes genes according to rhythmicity, zonation, and effect of TCDD treatment using a non-linear mixed effects model.

To investigate the spatial and zoned expression profiles of genes and assess how these profiles are altered by TCDD treatment, we adopted and extended the approach outlined in Droin et al²⁹. In their approach, rhythmicity (R) is modeled using sine and cosine functions, and zonation (Z) is modeled using first and second-order Legendre polynomials. We expanded on their model by including the effect of TCDD exposure (Dioxin; D) on gene expression across the peri-central axis of the liver lobule. We represented TCDD treatment in the model with a simple indicator function, where TCDD-treated hepatocytes were denoted as D=1, and untreated hepatocytes as D=0. Using

these variables, we constructed a series of non-linear mixed-effects models (NLMEMs) comprising different combinations of factors (D, Z, and R) that could potentially influence gene expression based on the experimental design and zonation inference (Fig. 24 and 25). The effects of these factors can be independent (e.g., $Z + R$) or dependent (e.g., $D \times Z$), as illustrated in Fig. 6. With this modeling approach, we classified the top 15,000 highly variable genes according to the best fitting NLMEM model to describe each gene's expression pattern (see methods). The NLMEM approach allowed us to model the complex interplay between TCDD exposure, rhythmic gene expression, and zonation patterns within the liver lobule. By incorporating the TCDD treatment factor (D) into the model, we could assess the impact of dioxin exposure on the spatial and temporal expression dynamics of individual genes. The model formulations included both independent and dependent effects, enabling the capture of various scenarios. For instance, the independent effects (e.g., $Z + R$) would describe genes whose zonation and rhythmicity patterns are not influenced by TCDD treatment, while the dependent effects (e.g., $D \times Z$) would capture genes whose zonation patterns are altered by TCDD exposure. By fitting these NLMEM models to the gene expression data and selecting the best-fitting model for each gene, we classify the top highly variable genes into different categories based on their expression patterns. This classification provides insights into the diverse transcriptional responses to TCDD treatment, including genes exhibiting altered zonation patterns, disrupted rhythmicity, or a combination of both.

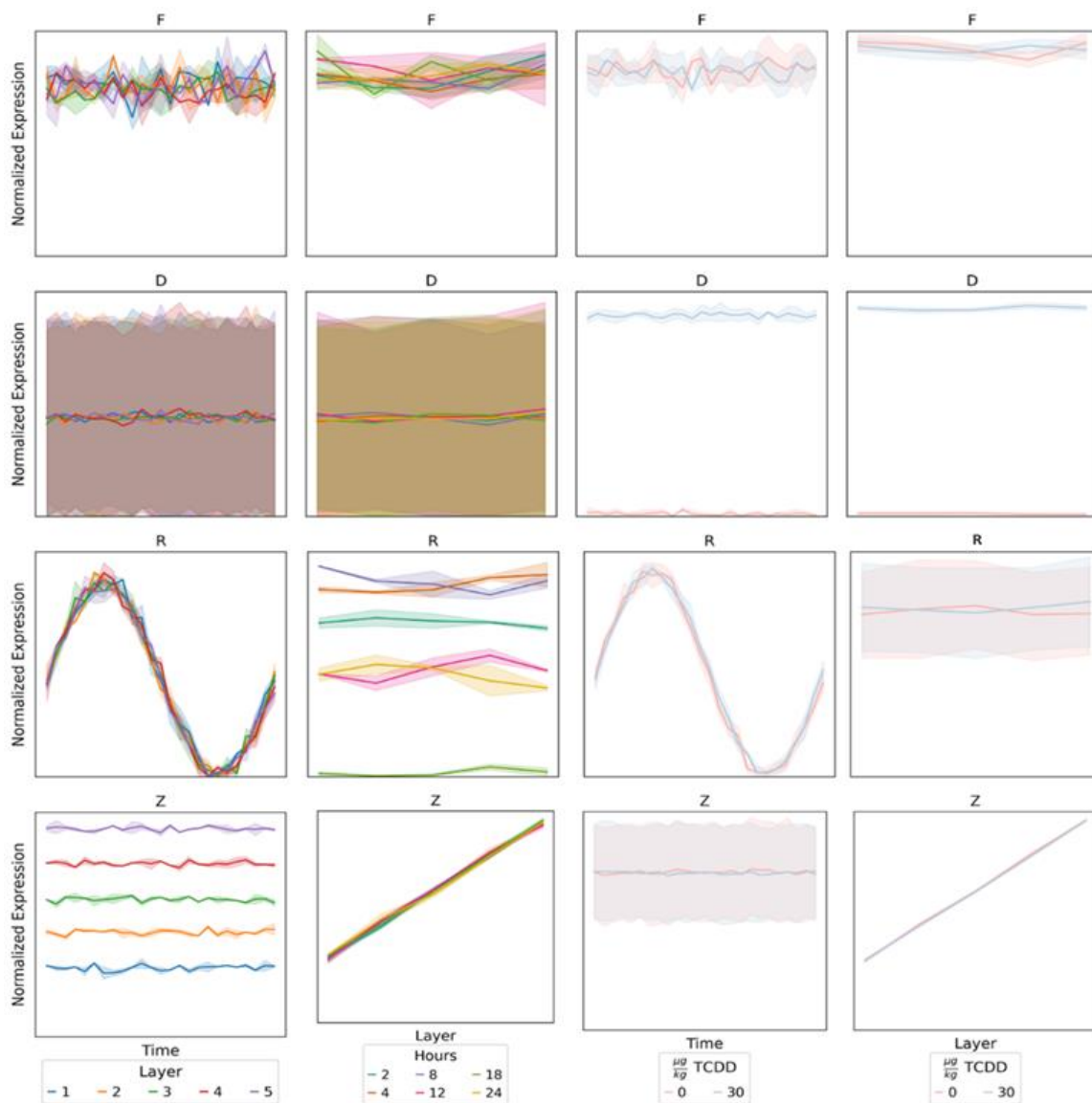


Figure 24: Schematic illustrations of each single or no effect non-linear mixed effect model profiles for classification of hepatocyte genes. Ideal simulations of each model plotted in terms of Time (in hours after treatment) versus expression or inferred layers of the liver lobule. Each line plot is 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.

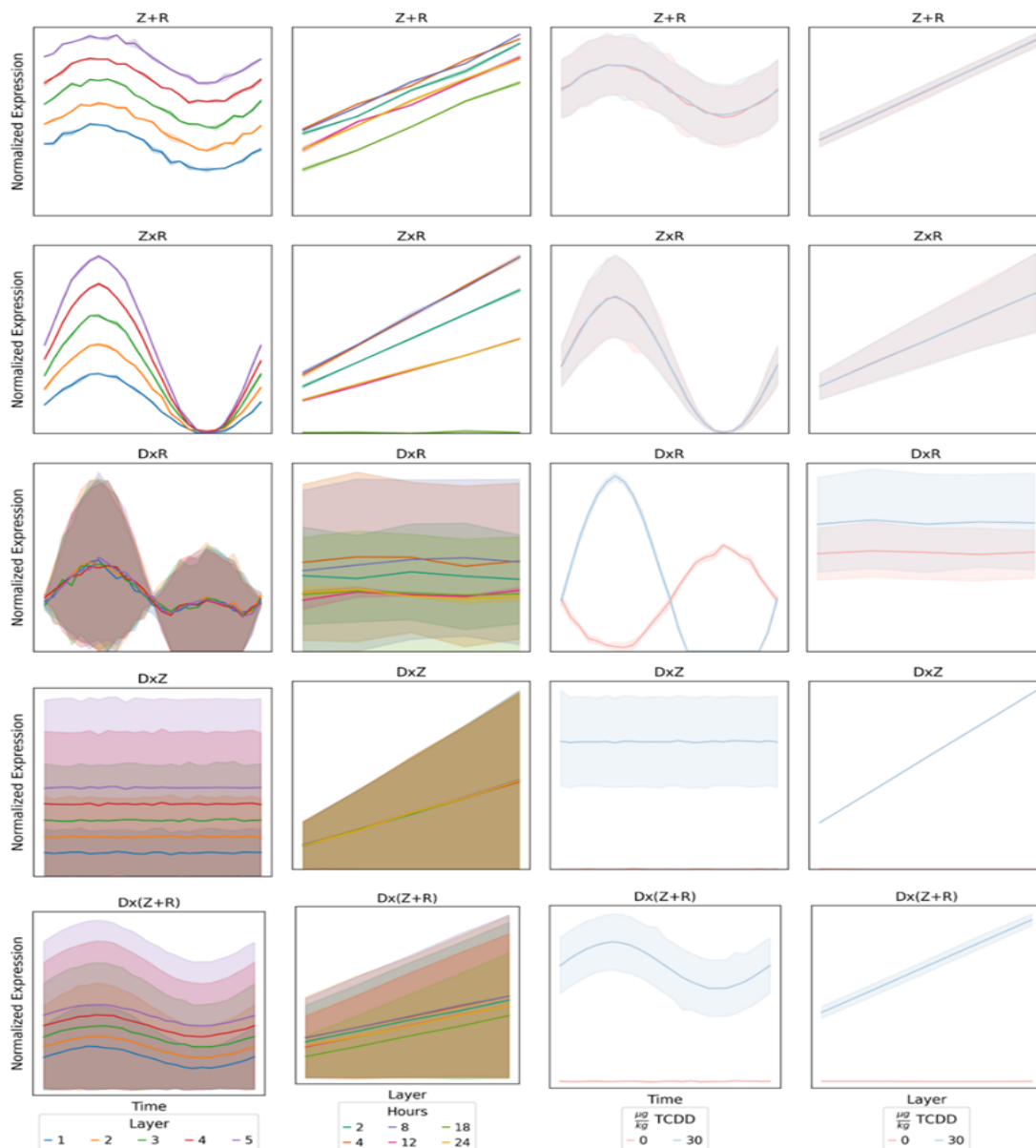


Figure 25: Schematic illustrations of multiple effect non-linear mixed effect model profiles for classification of hepatocyte genes. Ideal simulations of each multiple model category plotted in terms of Time (in hours after treatment) versus inferred layer of the liver lobule. Each line plot is colored either by time and layer of dose of TCDD of treatment. Effects that are 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. Example: Z+R indicates genes that are zonated and rhythmic independently and ZxR indicates genes that are zonated and rhythmic dependently.

To evaluate the accuracy of my gene classification approach, we initially examined a set of known genes with established rhythmic, zoned, or TCDD-responsive expression patterns. These included *Arntl* (known for rhythmic expression), *Slc1a2* (exhibiting both zoned and rhythmic patterns), and *Ahrr* (a gene induced by TCDD with saturation observed at 12 hours). As shown in Fig. 26, the model successfully captures the rhythmic profiles of *Arntl* and *Slc1a2*. The rhythmic expression patterns of these genes are accurately represented by the periodic functions incorporated in the model formulations. Additionally, the model can differentiate between TCDD saturation effects and true temporal rhythmicity. This distinction is crucial, as some genes may exhibit saturation in expression levels due to TCDD treatment, which should not be mistaken for rhythmic oscillations. Furthermore, the model correctly identifies the zoned pattern of *Slc1a2*. The incorporation of Legendre polynomials in the model allows for the accurate modeling of spatial zonation patterns within the liver lobule. Finally, *Ahrr* is properly classified by the model as being influenced by TCDD. The model captures the TCDD-induced expression changes in this gene, as evidenced by the distinct expression profiles between the treated and untreated conditions. Taken together, the expected expression patterns of these control genes are recapitulated, thereby validating the performance of our classification model. The model's ability to accurately capture rhythmic, zoned, and TCDD-responsive expression patterns in these well-characterized genes provides confidence in its applicability to the classification of the larger gene set.

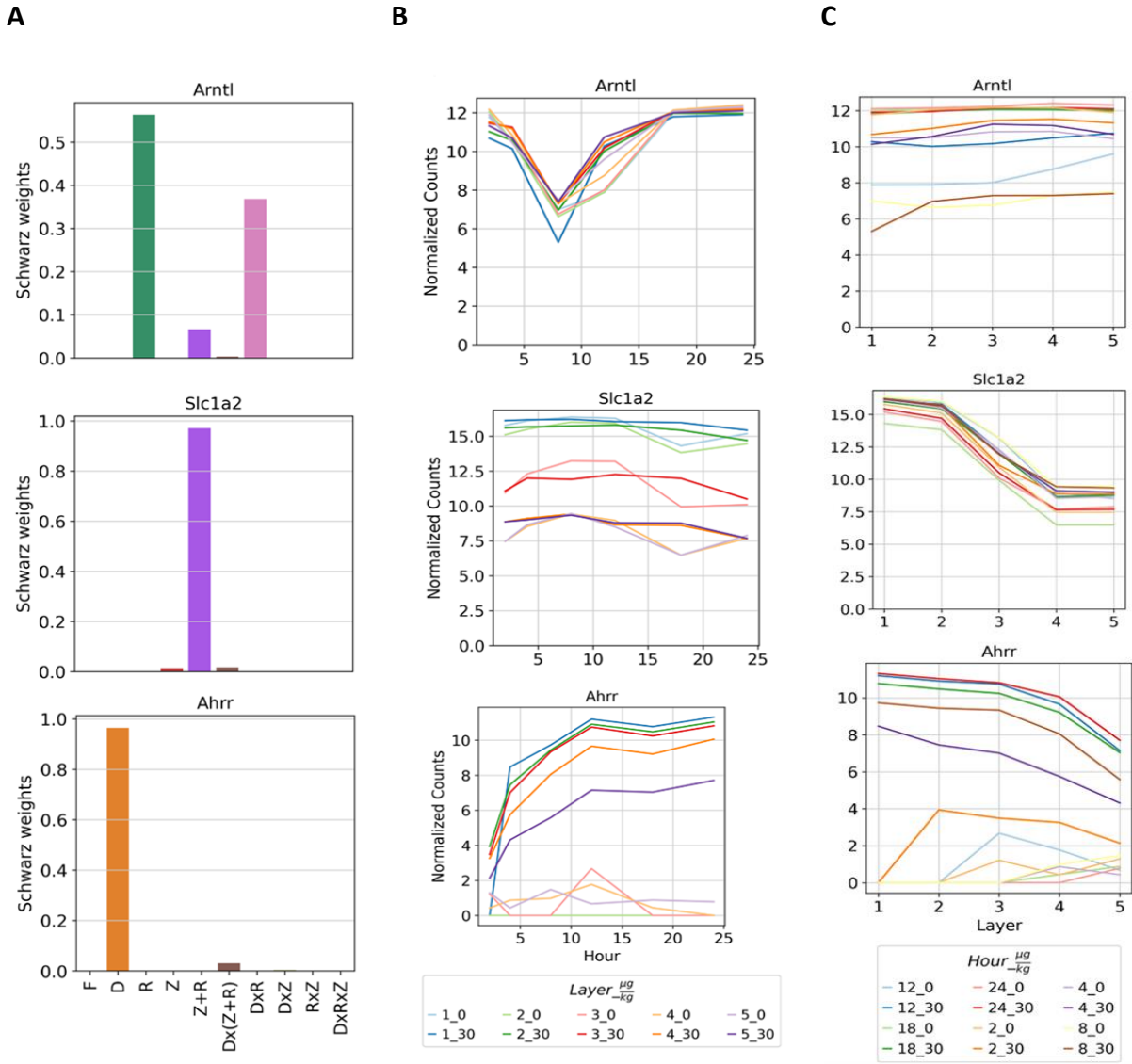


Figure 26: Classification of canonical rhythmic, zonal and TCDD responsive genes. (A) Bar plots of Schwarz weigh classification of Arntl, Slc1a2, and Ahrr. Line plots of expression colored by (B) Layer and TCDD dose. 1_0 denote lineplot expression of gene in the pericentral region after 0 $\mu\text{g/kg}$ dose treatment and 5_30 denote lineplot expression of gene in the periportal region after 30 $\mu\text{g/kg}$ dose treatment. (C) Hours after treatment and dose of TCDD. 2_0 denote lineplot expression of gene at timepoint 2hrs after 0 $\mu\text{g/kg}$ dose treatment and 24_30 denote lineplot expression of gene timepoint 24hrs after 30 $\mu\text{g/kg}$ dose treatment.

Finally, we categorized each gene mRNA profile into one of ten distinct patterns based on its expression dynamics: Flat (F) for profiles lacking rhythmic, zonated, or TCDD-induced influences; Rhythmicity (R) for genes exhibiting purely rhythmic expression patterns; Zonated (Z) for genes displaying solely zonated expression patterns; Dioxin (D) for genes induced by TCDD treatment; independent Zonated and Rhythmicity (Z+R) for genes showing both zonated and rhythmic expression patterns independently; dependent Zonated and Rhythmicity (Z×R) for genes with both zonated and rhythmic expression patterns dependently; dependent Dioxin and Rhythmicity (D×R) for genes exhibiting both TCDD-induced influence and rhythmic expression patterns dependently; dependent Dioxin and Zonated (D×Z) for genes displaying both zonated and TCDD-induced influence independently; independent Zonated and Rhythmicity with TCDD influence (D×(Z+R)) for genes showing both zonated and rhythmic expression patterns independently and influenced by TCDD treatment; and dependent Zonated and Rhythmicity with TCDD influence (D×(Z×R)) for genes exhibiting both zonated and rhythmic expression patterns dependently and influenced by TCDD treatment. The classification revealed that over 39% of the gene mRNA profiles were recorded as Flat (F), indicating no significant rhythmic, zonation, or TCDD influence. Additionally, 16.63% of the mRNA profiles were classified as Zonated (Z), while 15.15% and 8.84% were classified as Rhythmic (R) and Dioxin (D), respectively. These percentages agree with a previous study by Droin and Kholtei et al.¹, where they recorded more mRNA profiles classified as Zonated (Z) compared to the Rhythmicity (R), independent Zonated and Rhythmicity (Z+R), and dependent Zonated and Rhythmicity (ZxR) classes. The classification of gene expression profiles into these 10 patterns allowed for a comprehensive characterization of the diverse transcriptional responses to TCDD treatment, rhythmicity, and zonation within the liver lobule. By accounting for the independent and dependent effects of these factors, as well as their

potential interactions, we could gain insights into the complex regulatory mechanisms governing gene expression dynamics in hepatocytes. Furthermore, the agreement between our findings and those of the previous study by Droin et al²⁹. regarding the prevalence of zonated gene expression patterns provides additional confidence in the validity of our approach and the biological relevance of the observed patterns.

The classification also revealed that about 32% of the mRNA profiles exhibited rhythmic expression patterns, encompassing classes with a rhythmic component like R, Z+R, RxZ, DxR, etc. Additionally, approximately 35% of the mRNA profiles showed zonated expression patterns, including classes with a zonation component such as Z, Z+R, RxZ, DxZ, etc. Further analysis revealed that nearly 50% of rhythmic genes (mRNA profiles classified with an R component) were impacted by TCDD treatment, compared to around 25% of zonated genes (mRNA profiles classified with a Z component) (Fig. 27 B). This finding suggests that TCDD treatment had a more pronounced effect on disrupting rhythmic gene expression patterns compared to zonated patterns. The largest multi-effect category was $D \times (Z+R)$, which exhibited synergistic effects of TCDD on zonation and rhythmicity independently (Fig. 27 A). This class represents genes whose zonation and rhythmic expression patterns were independently influenced by TCDD treatment. Gene set enrichment analysis of the $D \times (Z+R)$ class showed that these genes were enriched for canonical pathways involved in TCDD liver toxicity, such as metabolism of xenobiotics and drug metabolism by cytochrome P450 (Fig. 27 C). This enrichment indicates that the genes in this class play crucial roles in the liver's response to TCDD exposure, potentially contributing to the observed toxicological effects. Notably, no genes displayed dependence between dose, zonation, and rhythmicity ($D \times Z \times R$), suggesting that the effects of TCDD treatment on zonation and rhythmicity were independent in the context of this study. Overall, our results demonstrate that gene expression

in the liver is not static but is partitioned across the liver lobule in space (zonation), time (rhythmicity), and by treatment dose. The interplay between these factors results in a complex landscape of gene expression patterns, with TCDD treatment exerting differential effects on rhythmic and zoned gene expression profiles. By integrating spatial, temporal, and treatment-related factors in our analysis, we could comprehensively characterize the transcriptional dynamics within the liver lobule and elucidate the impact of TCDD exposure on these intricate regulatory mechanisms. This approach provides a more holistic understanding of the molecular processes governing liver function and toxicological responses, contributing to the development of targeted interventions and therapeutic strategies.

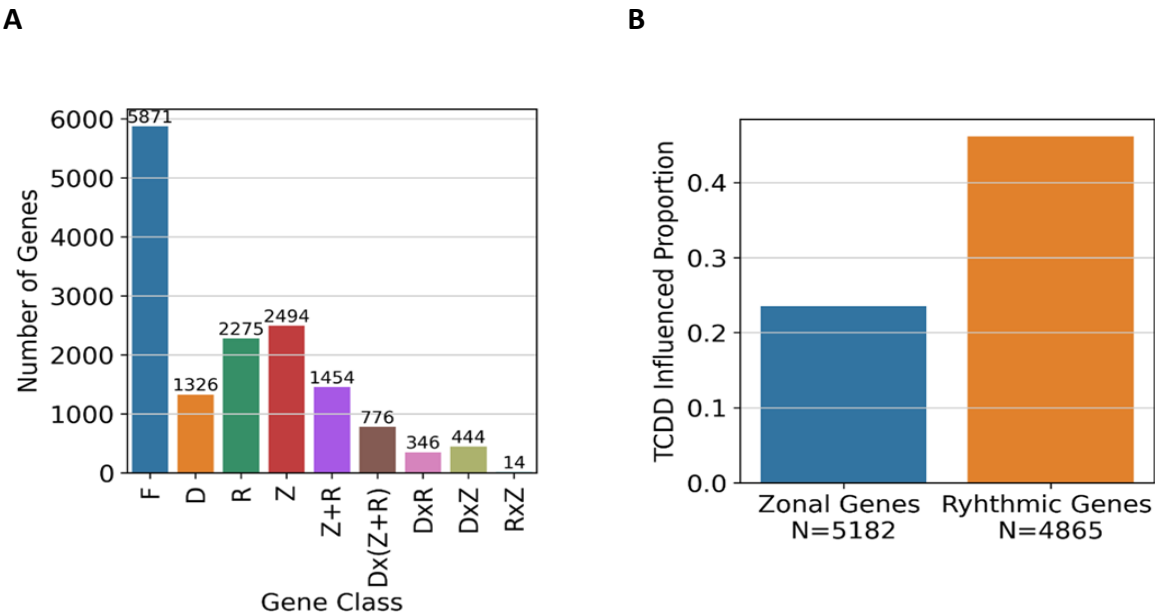
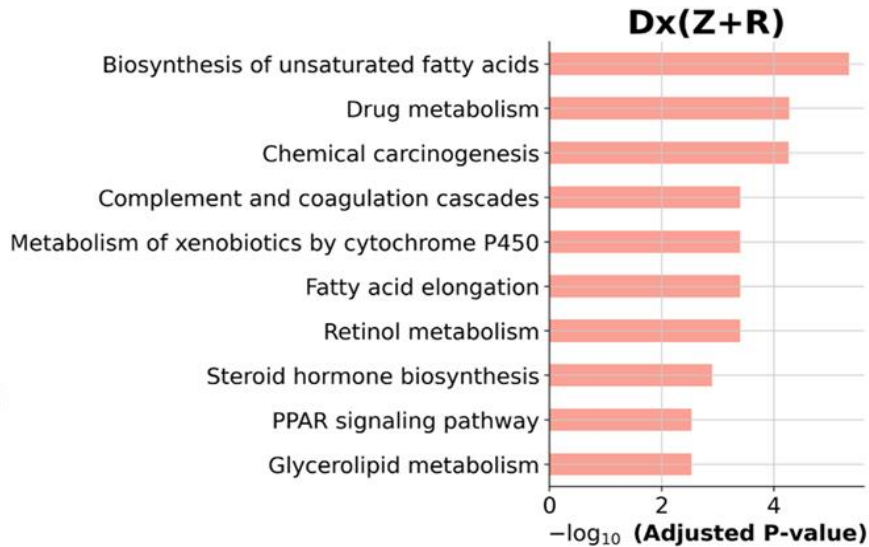


Figure 27: Classification distributions and enrichment of TCDD toxicity pathways. (A) Bar plot of the number of genes (mRNA profiles) in each class for the top 15,000 HVGs. (B) Bar plot of the proportion of genes that have TCDD influence for all zonal genes (5182 total zonal genes) and all rhythmic genes (4865 total rhythmic genes). (C) GSEA analysis of the Dx(Z+R) class of genes.

Figure 27 (cont'd)

C



The effects of acute 2,3,7,8 Tetrachlorodibenzo-p-dioxin (TCDD) perturbation of rhythmic genes.

Utilizing the classification framework illustrated in Figure 27A, our investigation delved into the acute toxicity effects on rhythmic genes within the liver lobule. We first analyzed the the impact of TCDD on a core set of circadian rhythm genes (Clock, Arntl, Per2, Cry1, Nr1d1, Npas2, Rorc) by by examining their mRNA transcript classification. Notably, we found that all core circadian clock genes exhibited rhythmicity in their classification classes. Only Arntl, Clock and Rorc were classified as purely rhythmic (Fig. 28 A). However, upon closer examination of their expression patterns, we found that Arntl and Clock exhibited significantly increased expression between 11 and 16 hours ($p\text{-value} < 0.01$, Mann-Whitney U-test) with TCDD treatment. This observation aligns with the saturation of TCDD response genes analyzed in Figure 20 F. Conversely, Npas2, Per2, and Nr1d2 were classified as having TCDD influence (Fig 28 A). Per2 had significantly higher expression for all time points except 18 hours post-treatment. Npas2 had significantly lower

expressions at timepoints 2, 4, and 24. Nr1d2 exhibited significantly higher expression at timepoints 2 and 4, but significantly lower expression at timepoint 12. Cry1 was classified as zonal and rhythmic, which corroborates previous studies. However, the second-highest classification of Cry1 implied TCDD influence, observable at timepoints 18 and 24.

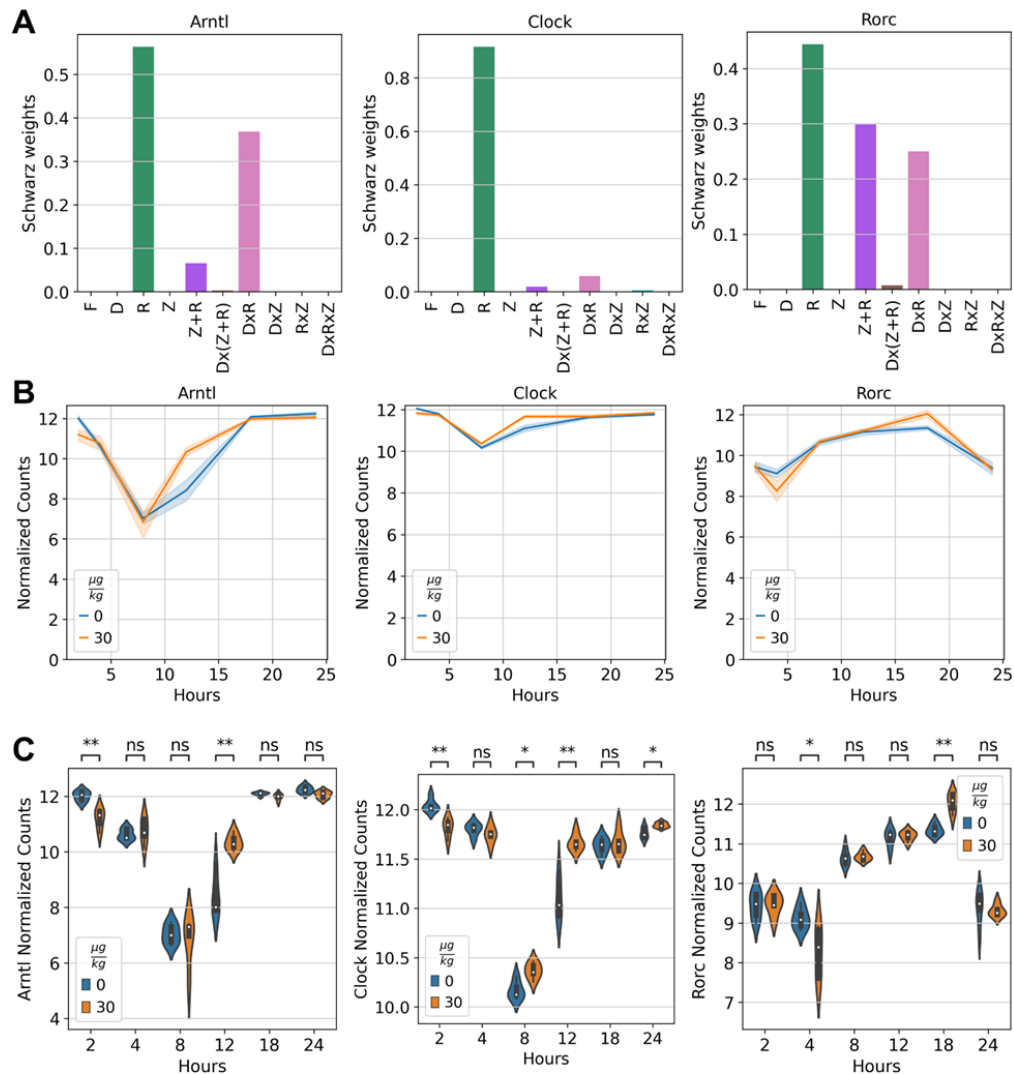


Figure 28: Core circadian clock genes classified as only rhythmic. (A) Bar plots of Schwarz weights for gene classification. (B) Line plots of gene expression graphed with respect to hours after treatment and colored by dose of treatment of TCDD. (C) Violin plots plotting the distribution of expression for each treatment condition at each timepoint.

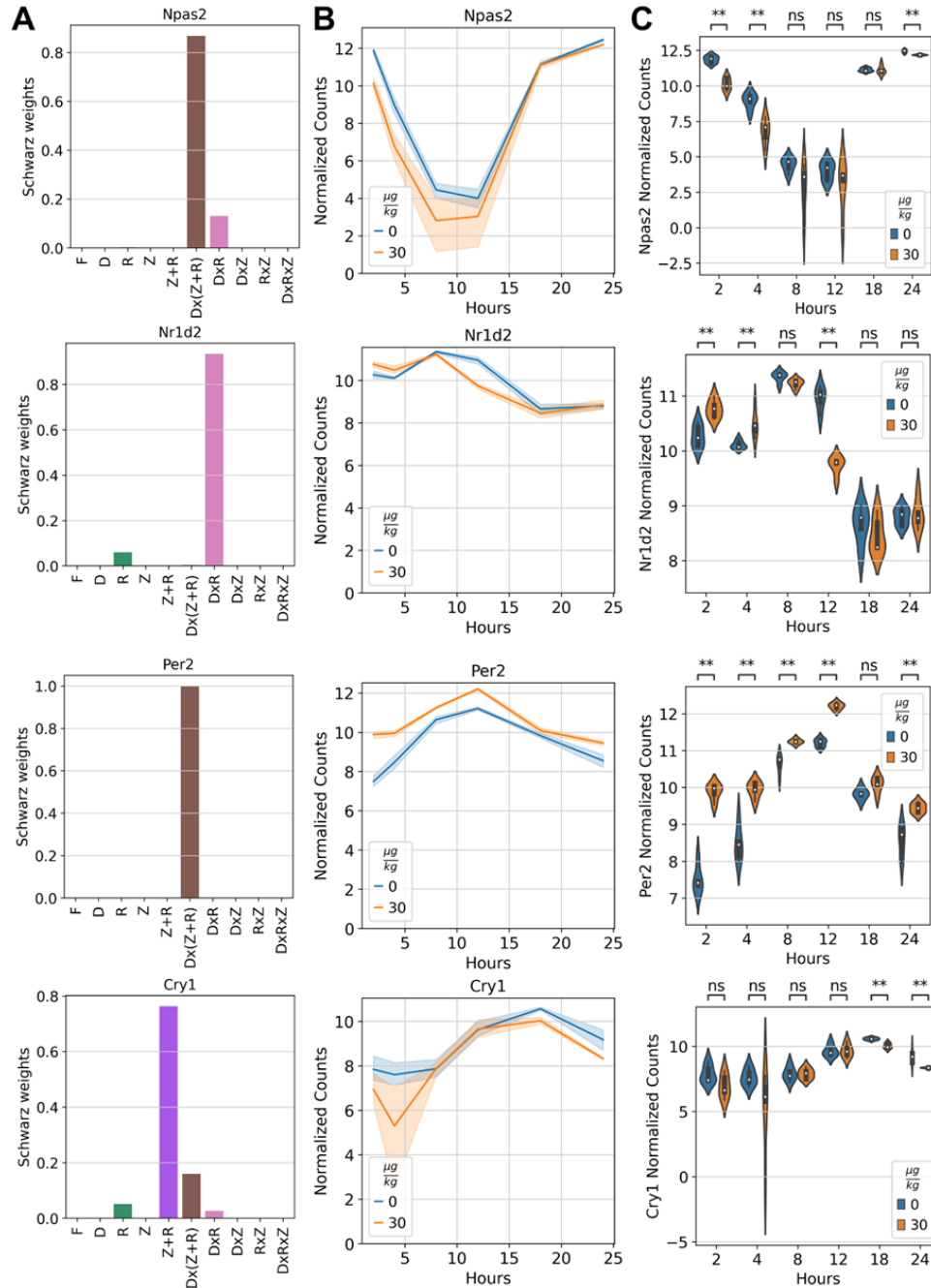
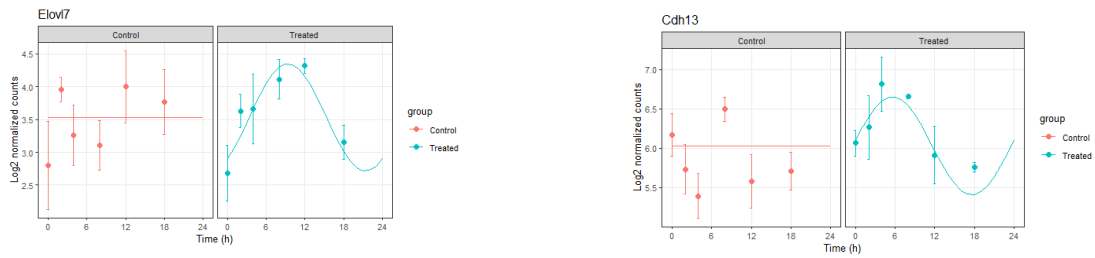


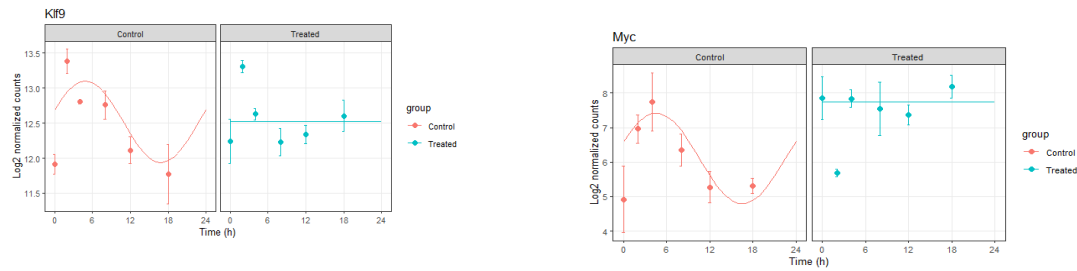
Figure 29: Core circadian clock genes classified as having multiple effects. (A) Bar plots of Schwarz weights for gene classification. (B) Line plots of gene expression graphed with respect to hours after treatment and colored by dose of treatment of TCDD. (C) Violin plots plotting the distribution of expression for each treatment condition at each timepoint.

To quantify the effects of TCDD on rhythmic gene expression, we explored whether TCDD treatment induced or removed rhythmicity in genes classified as rhythmic and influenced by TCDD (DxR and Dx(R+Z)). First, we used dryR (Differential Rhythmicity analysis in R)^{132,137}, a package in R, to assess the differential rhythmicity of the control and treated genes found in the DxR and Dx(R+Z) classes. This analysis unveiled that genes such as *Elovl7* and *Cdh13* gained rhythmicity subsequent to TCDD exposure, whereas *Klf9* and *Myc* lost their rhythmic patterns (as depicted in Fig 29 A and B). We further analyzed rhythmicity changes by using a likelihood ratio test to compare whether gene expression better fit a rhythmic or flat (non-rhythmic) model. Genes fitting a rhythmic model under control but a flat model under treatment were deemed to have lost rhythmicity. Conversely, those fitting a flat expression model in control but a rhythmic model in treatment were classified as having gained rhythmicity. Among TCDD-influenced rhythmic genes, distribution analysis revealed 13% lost rhythmicity after treatment, while 21% gained rhythmicity (Fig 29 C). Subsequent gene set enrichment analysis (GSEA) conducted on both the gained and lost rhythmicity groups elucidated hallmark pathways of TCDD toxicity. The lost rhythmicity group showed enrichment for processes related to chemical carcinogenesis, drug metabolism and metabolism of xenobiotics by cytochrome P450, while the gained rhythmicity group was enriched for retinol metabolism¹³⁸. In summary, TCDD exposure induced both gains and losses of rhythmicity in different sets of genes, with the affected genes enriched for processes related to TCDD toxicity and metabolism.

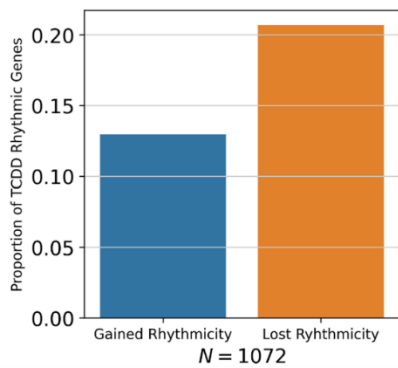
A



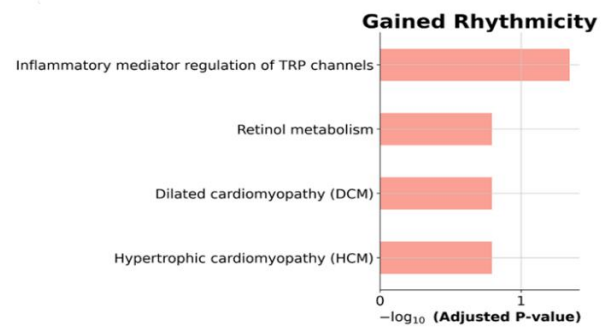
B



C



D



E

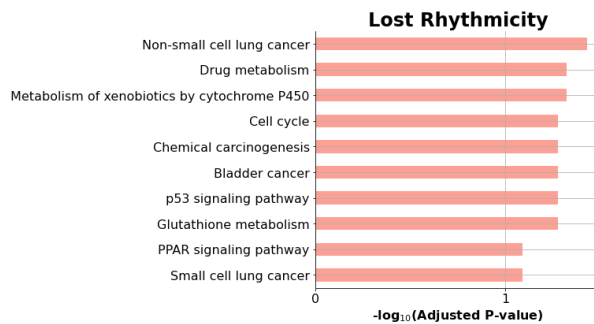


Figure 30: Analysis of gain or loss of rhythmicity for TCDD influenced rhythmic genes.

Figure 30 (cont'd)

Time series expression profile (A) gained rhythmicity genes *Elov17*, and *Cdh13* (B) lost rhythmicity genes *Klf9*, and *Myc* for control and treated set. (C) 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 using dryR. A bar plot of the proportion of genes analyzed colored by whether they lost or gained rhythmicity. (D) GSEA analysis of gene sets for genes that gained rhythmicity. (E) GSEA analysis of gene sets for genes that lost rhythmicity.

Further analyzing genes that maintained rhythmicity after TCDD exposure, we investigated how TCDD impacted the properties of their rhythmic expression patterns, specifically the amplitudes and phases. When analyzing the core circadian clock genes, we observed only small changes in the phase and amplitude of these genes (Figure 30 B) after TCDD treatment. Extending this analysis to all genes that kept rhythmicity, we similarly found no major trends reflecting a delay in phase or reduction of amplitude after TCDD exposure (Figure 30 C). To determine if the magnitude of gene expression changes induced by TCDD treatment correlated with changes in the phase or amplitude of rhythmic expression, we examined the relationship between these parameters. However, we found only a weak correlation between the magnitude of gene expression changes and the magnitude of phase or amplitude changes (Figure 30 D). These findings suggest that while TCDD exposure induced gains and losses of rhythmicity in certain gene sets, the genes that maintained rhythmicity after treatment generally exhibited minimal alterations in their rhythmic properties, such as phase and amplitude. Furthermore, the extent of gene expression changes induced by TCDD did not strongly correlate with the degree of phase or amplitude shifts in the rhythmic expression patterns of these genes.

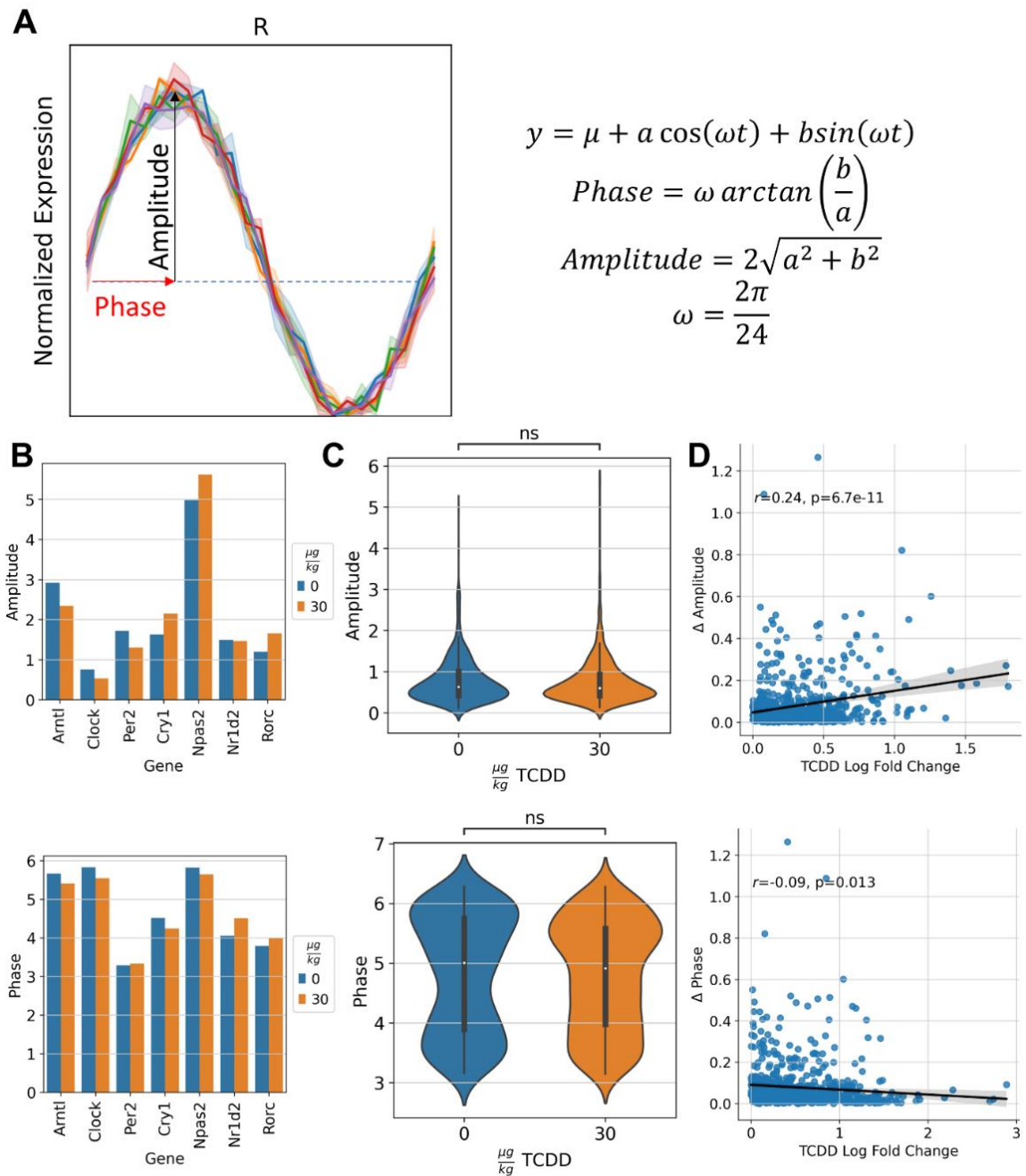


Figure 31: Analysis of rhythmic parameters for TCDD influenced rhythmic genes. (A) Schematic and equations describing phase and amplitude of gene's expression. (B) Bar plot of core circadian clock genes and their respective phase and amplitude in each treatment group. Phase is measured in radians, and 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.

The effects of acute 2,3,7,8 Tetrachlorodibenzo-p-dioxin (TCDD) perturbation of rhythmic genes.

To elucidate the effects of TCDD on zonal gene expression patterns in the liver, we examined the enrichment of known transcriptional targets of the Wnt/ β -catenin, Ras, and hypoxia signaling pathways, which are primarily responsible for regulating hepatic zonation^{35,139}. Our investigation aimed to ascertain whether each zoned gene class exhibited enrichment in these zonation pathways, considering the presence or absence of TCDD influence. Our analysis unveiled that all targets associated with zonation pathways demonstrated enrichment within the Z+R class of genes, which exhibited both zonal and rhythmic expression patterns. Interestingly, the TCDD-perturbed Dx(Z+R) genes and the dual-effect DxZ class, which exhibited both TCDD influence and zonal expression, were enriched for Wnt and hypoxia pathway target genes but not for Ras pathway target genes (Figure 31). These results demonstrate that TCDD selectively disrupts Ras signaling-mediated periportal gene expression, while the zonation programs regulated by the Wnt/ β -catenin and hypoxia signaling pathways remain largely intact. In summary, our findings suggest that TCDD exposure selectively perturbs the Ras signaling pathway, which is responsible for regulating the expression of periportal genes, while leaving the Wnt/ β -catenin and hypoxia-mediated zonation programs relatively unaffected. This selective disruption of Ras signaling may contribute to the hepatotoxic effects of TCDD and highlights the complex interplay between xenobiotic exposure, signaling pathways, and spatial gene expression patterns in the liver.

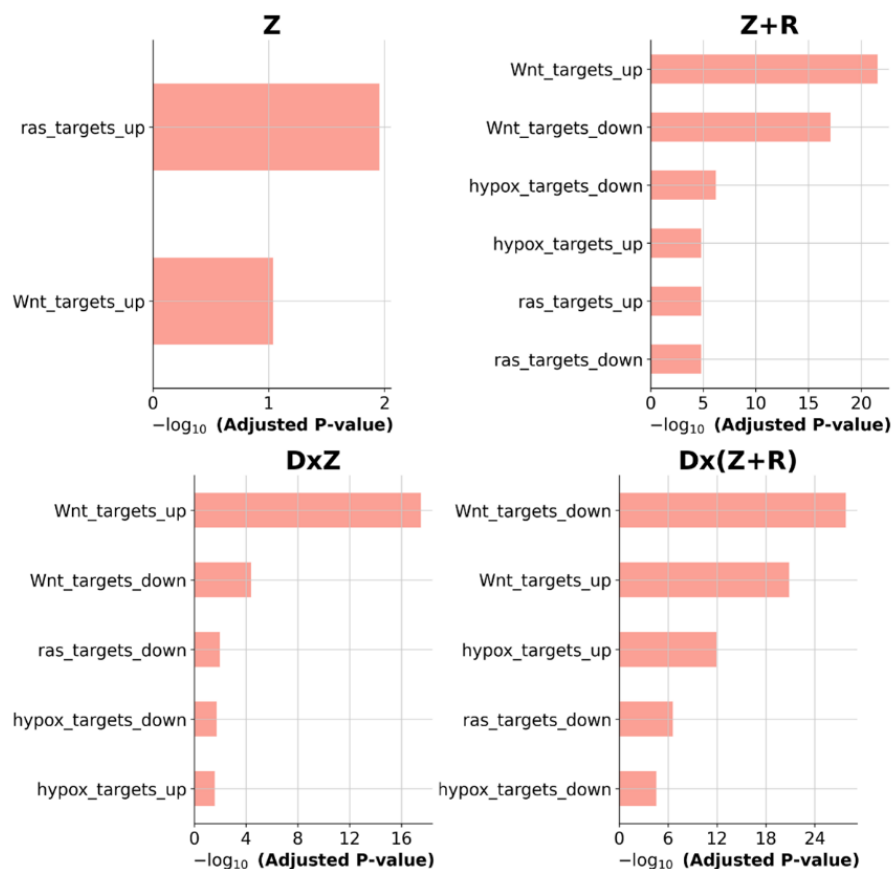


Figure 32: Enrichment analysis of zonation regulation pathway targets on all zonation gene classes. Building upon the analysis of rhythmic gene expression, we conducted an analogous investigation of differential zonation to discern whether genes in the DxZ and Dx(Z+R) categories experienced alterations in their zonal expression patterns following TCDD treatment. Using a likelihood ratio test, we performed zonation differential expression analysis to identify genes that gained, lost, or maintained zonal specificity with TCDD exposure. Our findings indicate that 18% of the analyzed genes gained zonation after TCDD treatment, while 13% lost their zonal expression patterns (Figure 32). Gene set enrichment analysis revealed that pathways related to UDP-glucuronosyltransferase enzymes, such as "Pentose and Glucuronate Interconversions," were overrepresented among genes losing zonation (Figure 32), with no pathways enriched in the gained

zonation set. Furthermore, in analyzing whether the genes that lost or gained zonation exhibited a more portal or central enrichment, we observed that genes losing zonation were significantly more centrally enriched compared to the background (p-value < 0.001, Kolmogorov-Smirnov Test), indicating selective disruption of periportal gene expression by TCDD (Figure 32). Together, these results demonstrate that TCDD exposure leads to bidirectional alterations in hepatic zonation, with apparent centrilobular targeting of UDP-glucuronosyltransferase zonal discontinuity. The selective disruption of periportal gene expression patterns and the enrichment of UDP-glucuronosyltransferase pathways among genes losing zonation suggest a potential mechanism for TCDD-induced hepatotoxicity through impaired xenobiotic metabolism and clearance.

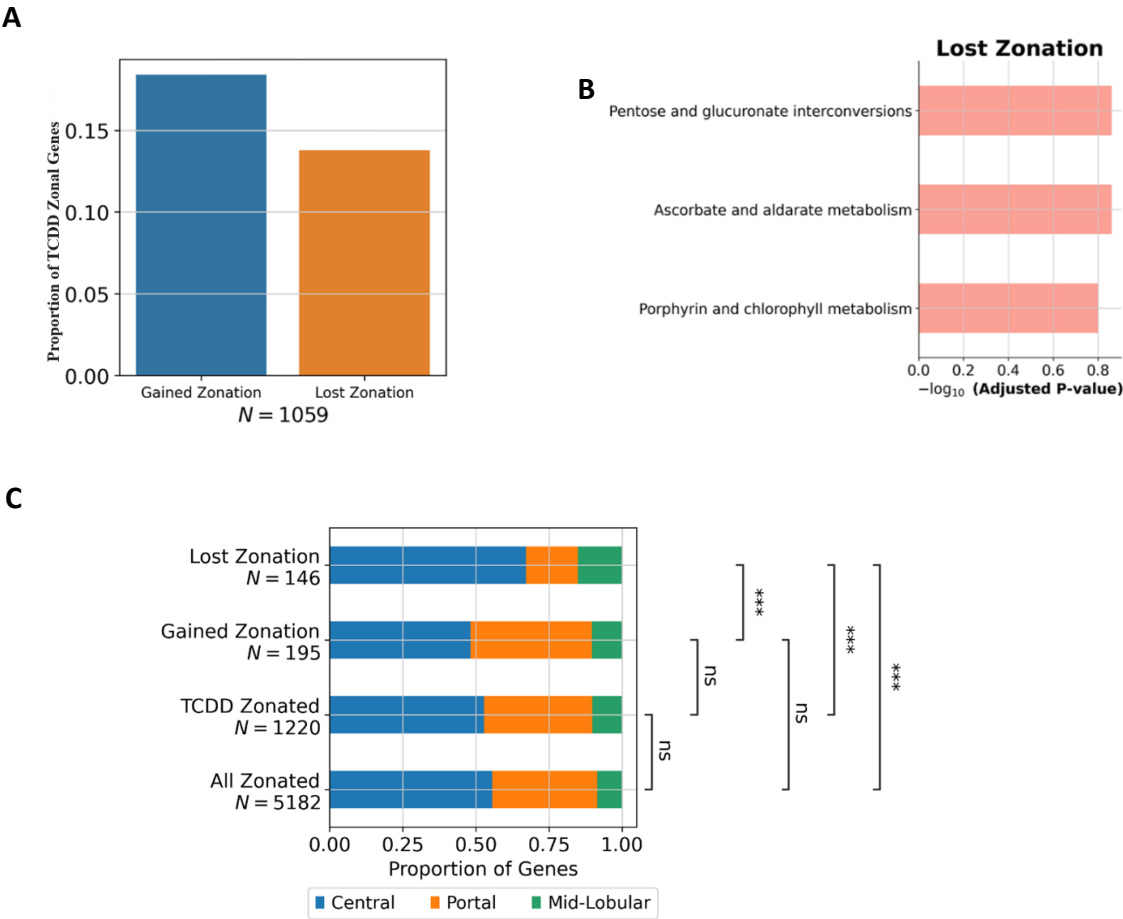


Figure 33: Analysis of loss or gain of zonation based on TCDD treatment. (A) A total of 1059 genes that were classified to be influenced by Zonation and TCDD were analyzed as to whether

Figure 33 (cont'd)

they had gained, or lost zonation based on the likelihood ratio. (B) GSEA analysis of genes that lost zonation with TCDD treatment. (C) Stacked bar plot describing the distribution of zoned genes and in which zone of the liver lobule are those genes most highly expressed. Differences in distribution calculated using the Kolmogorov-Smirnov Test. ** means significant difference between the two expressions at the time point and ns means no significant difference between the two expressions.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) analysis of dose affected categories.

To understand the transcriptional regulation across various gene class categories under the influence of TCDD, which impacts both zonal and rhythmic gene expression, we investigated the presence of binding motifs for the transcription factors BMAL1⁷⁵ (E-box motifs) and AHR^{75,140} (Dioxin Response Elements, DREs) in these genes. While the presence of motifs is necessary but not sufficient for transcription factor binding, we focused specifically on motifs located in accessible chromatin regions, mapped by DNase I hypersensitivity (DHS) data in mouse liver tissue from the ENCODE project. Using the GRCm38 reference genome and BEDTools⁶⁸, we extracted all canonical E-box (CANNTG) and DRE (GCGTG) motifs located in DHS peaks. We then matched these accessible motifs to hepatic AHR ChIP-seq data in male C57BL/6 mice treated with 30 µg/kg TCDD, to identify motifs likely bound by transcription factors. Our approach involved extracting genes with E-box and DRE binding motifs in their promoter or genome region, as well as identifying genes with an overlap of both E-box and DRE motifs (E-box intersect DREs). By integrating this information with our gene classification data, we aimed to elucidate the potential transcriptional regulatory mechanisms governing the observed gene expression patterns under TCDD exposure. Specifically, we investigated whether the presence of accessible E-box, DRE, or overlapping E-box/DRE motifs could explain the differential expression, rhythmicity, and zonation patterns observed in various gene classes. By benchmarking these gene sets against the

classification classes derived from our non-linear mixed effects models (NLMEMs), we uncovered intriguing insights. Specifically, more than 55% of genes in the Dx(R+Z) class, which exhibited both TCDD influence and combined zonal and rhythmic expression patterns, contained DRE (Dioxin Response Element) binding motifs, 20% exhibited E-box binding motifs, and over 5% possessed overlapping E-box and DRE binding motifs (E-box intersect DRE) (Figure 33 A). Gene set enrichment analysis further illuminated distinct pathways associated with these binding motifs. For genes enriched with only E-box binding motifs in the Dx(R+Z) class, pathways related to chemical carcinogenesis, drug metabolism, and xenobiotic metabolism by cytochrome P450 were prominent (Figure 33 B). Meanwhile, genes featuring only DRE binding motifs in the Dx(R+Z) class were enriched in pathways like the PPAR signaling pathway, p53 signaling pathway, and glycerolipid metabolism (Figure 33 C). Notably, genes with overlapping E-box and DRE binding motifs (E-box intersect DRE) in the Dx(R+Z) class exhibited enrichment in pathways related to vitamin B6 metabolism and circadian rhythm (Figure 33 D). These findings suggest that the interplay between the circadian clock machinery (E-box motifs) and the xenobiotic response pathway (DRE motifs) plays a crucial role in regulating hepatic gene expression under TCDD exposure. The presence of accessible E-box and DRE motifs in specific gene classes may explain the observed differential expression, rhythmicity, and zonation patterns. Furthermore, the distinct pathway enrichments associated with genes containing E-box, DRE, or overlapping E-box/DRE motifs provide insights into the biological processes potentially regulated by these transcription factor binding sites under TCDD exposure.

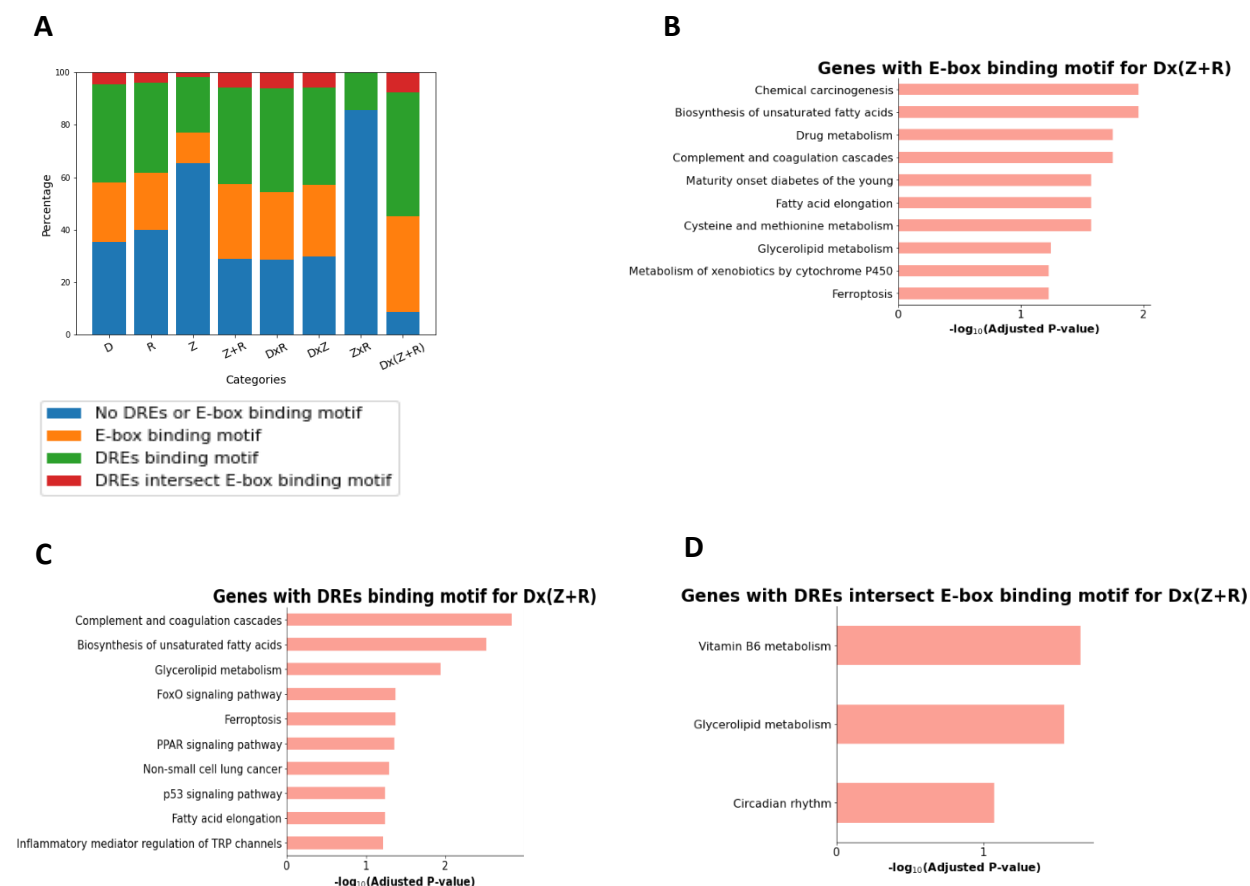
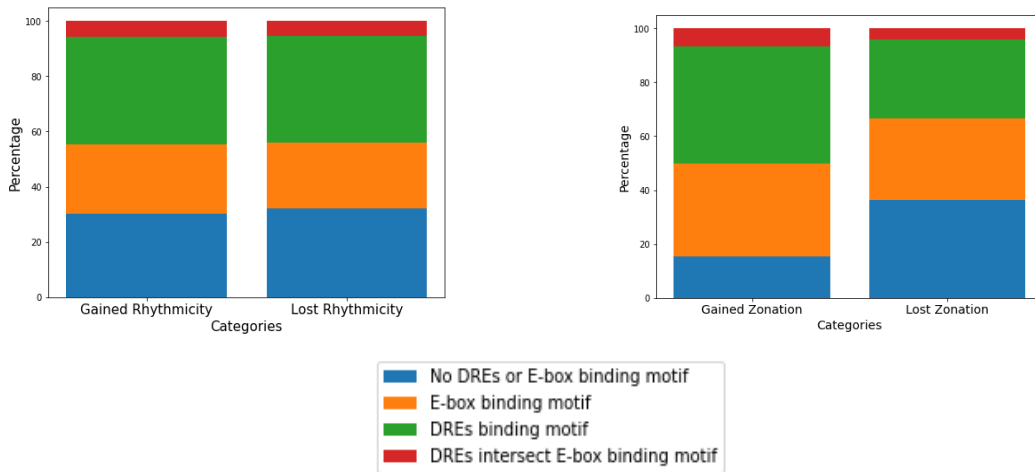


Figure 34: ChIP-Seq and GSEA analysis of classification categories. (A) ChIP-Seq analysis of the gene classification categories. Stacked histogram of genes enriched for BMAL1 binding (orange), genes enriched for AHR binding (green), genes enriched for AHR and BMAL1 binding (red) and genes enriched for neither AHR nor BMAL1 binding (blue). Gene-set enrichment analysis on (B) Genes enriched with BMAL1 binding (C) Genes enriched with AHR binding. (D) Genes enriched with AHR and BMAL1 binding.

We took a similar approach to characterize the presence of E-box motifs, DRE motifs, and overlapping E-box/DRE motifs in genes that gained or lost rhythmicity following TCDD treatment. Our analysis found that comparable numbers of rhythmicity-perturbed genes contained only E-boxes, only DREs, or intersecting E-box/DRE motifs. Over 50% of TCDD-affected zoned genes that lost spatial zonation contained DRE motifs in accessible promoter or genomic

regions. Notably, about 38% of arrhythmic genes lacked both DRE and E-box motifs, suggesting other mechanisms of regulation (Figure 34 A). Gene set enrichment analysis linked E-box/DRE-containing genes that lost zonation to pathways including starch/sucrose metabolism, fatty acid elongation, and proteoglycans in cancer. E-box/DRE motifs were also enriched in arrhythmic genes associated with bladder cancer, insulin signaling, FoxO signaling, and related pathways (Figure 34 B). No pathways were significantly enriched among gained rhythmicity genes based on our cutoff criteria. Overall, intersecting E-box and DRE motifs emerged as key potential mediators of TCDD-disrupted hepatic rhythmicity and zonation, though additional factors appear contributory for approximately 38% of rhythmicity-lost genes that lacked these motifs. These findings suggest that the interplay between the circadian clock machinery (E-box motifs) and the xenobiotic response pathway (DRE motifs) plays a crucial role in mediating the effects of TCDD on hepatic gene expression rhythmicity and zonation. The presence of accessible E-box and DRE motifs in specific gene sets may explain the observed gains or losses of rhythmicity and zonation following TCDD exposure. Furthermore, the distinct pathway enrichments associated with genes containing E-box, DRE, or overlapping E-box/DRE motifs provide insights into the biological processes potentially dysregulated by these transcription factor binding sites under TCDD exposure. However, for a subset of rhythmicity-lost genes (~38%), the lack of these motifs suggests the involvement of other regulatory mechanisms in mediating TCDD-induced arrhythmicity.

A



B

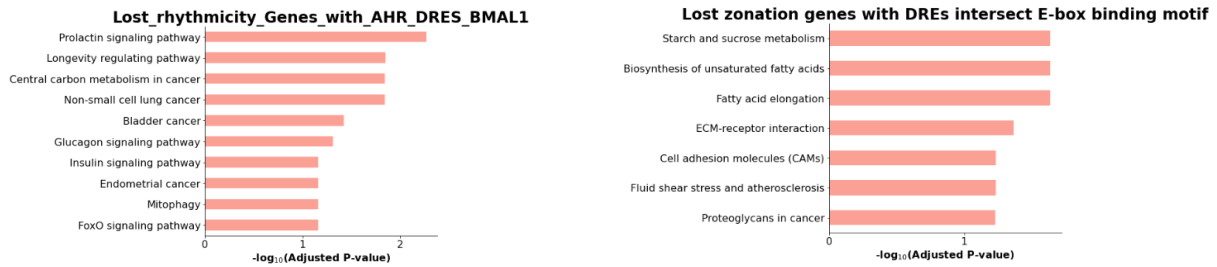


Figure 35: ChIP-Seq and GSEA analysis of TCDD perturbed rhythmic and zonated genes. (A) ChIP-Seq analysis of the TCDD perturbed rhythmic genes (gained and lost rhythmicity genes) and TCDD perturbed zonated genes (gained and lost zonation genes). Genes enriched for BMAL1 binding (orange), genes enriched for AHR binding (green), genes enriched for AHR and BMAL1 binding (red) and genes enriched for neither AHR nor BMAL1 binding (blue). (B) Gene-set enrichment analysis on lost rhythmicity and zonated genes with AHR and BMAL1 binding.

Comparison of Gene regulatory networks (GRNs) of control and treated data.

Gene regulatory networks (GRNs) are complex systems that govern gene expression patterns and cellular processes, influencing cellular functions and phenotypic outcomes. Traditional methods for inferring GRNs relies on bulk RNA-seq and chromatin immunoprecipitation sequencing (ChIP-seq) average out cellular heterogeneity and lack the resolution to capture regulatory dynamics at the single-cell level^{141,142}. Recent advances in single-cell sequencing technologies,

such as single-cell RNA sequencing (scRNA-seq) and single-nuclei RNA sequencing (snRNA-seq) have enabled the measurement of gene expression levels at an unprecedented resolution, capturing the heterogeneity and dynamics of individual cells within a population^{126,143,144} To understand the gene regulatory mechanism across the portal-central axis of the liver lobule and how this mechanism is impacted by TCDD, we employ ScGeneRAI¹⁴⁵. ScGeneRAI is an interpretable machine learning method that employs layer-wise relevance propagation (LRP) to infer GRNs from single-nucleus RNA sequencing data of cells in each layer. ScGeneRAI train a deep neural network to predict the expression of a target gene based on the expression of arbitrary sets of regulator genes in single cells. Subsequently, LRP quantitatively assigns each regulator gene a relevance score for the target gene expression prediction. Thereby, LRP identifies key transcription factors and constructs cell type specific GRNs de novo. Comparison of the layer specific GRNs generated by scGeneRAI then allowed us to determine conserved and distinct regulatory interactions in central versus portal hepatocytes. Utilizing the computational framework tailored for inferring GRNs from single-cell data, we unraveled the transcriptional regulation between transcription factors (TFs) and targeted genes along the portal-central axis of the liver lobule and the impact of TCDD on this mechanism. We analyzed the gene regulatory networks (GRNs) of central cells in layer 1 to portal cells in layer 5 to characterize transcription factor interactions with targeted genes and assess similarities and differences across the liver lobule. We compared the network similarities and differences across layers and across treatments (Control vs Treated).

Our analysis of single-cell transcriptomic data from liver cells revealed intriguing insights into the gene regulatory networks (GRNs) governing circadian rhythms and zonal gene expression patterns along the portal-central axis of the liver lobule. Employing the state-of-the-art ScGeneRAI

framework, we inferred GRNs from both control and TCDD-treated datasets, unveiling the regulatory interactions between transcription factors (TFs) and their target genes. In the control dataset, our findings highlighted a crucial interaction between the TF NPAS2 and the gene *Arntl*, which together form the master regulator of the circadian clock mechanism [2]. Notably, this interaction between NPAS2 and *Arntl* was the only observed interaction involving the NPAS2 TF, underscoring its importance in maintaining circadian rhythmicity. Moreover, our analysis revealed that several TFs, including FOXO3, PROX1, RXRA, and GPAM, exhibited substantial interactions with the *Bach2* gene, suggesting its potential involvement in various regulatory processes within liver cells. Intriguingly, upon analyzing the TCDD-treated dataset, we observed a striking disruption of the interaction between the NPAS2 TF and its target gene *Arntl*. This disruption was consistent across all layers of the liver lobule, implying a potential dysregulation of the circadian clock mechanism following TCDD treatment. Furthermore, our findings unveiled a gradual reduction in the number of interactions between TFs and target genes as we traversed from the central to the portal region of the liver lobule. This observation aligns with the well-established concept of zonal gene expression and metabolic function zonation within the liver lobule. Cells in the central lobule exhibited a higher density of regulatory interactions compared to those in the portal lobule, suggesting a more intense transcriptional regulation in the central region.

Finally, we investigated the similarities between the gene regulatory network (GRN) graphs within the control and treated datasets. To achieve this, we utilized the average number of interactions based on transcription factors, their target genes, and the interactions between them. The analysis involved comparing the GRN similarities across different layers of the liver lobule. Our findings revealed that the GRN similarities decrease across the liver lobule, indicating a gradual divergence

in gene regulatory networks as we move from the periportal zone to the pericentral zone. Specifically, we observed that the similarity between the GRNs of adjacent layers, such as Zone 1 and Zone 2 (Z1Z2), was higher than the similarity between the GRNs of more distant layers, such as Zone 1 and Zone 5 (Z1Z5). This observation suggests that the gene regulatory mechanisms governing cellular processes may differ across the liver lobule, potentially reflecting the varying metabolic demands and functional specializations of hepatocytes in different zones. The higher similarity between adjacent layers could be attributed to the gradual transition in cellular environments and the shared regulatory mechanisms between neighboring zones. It is important to note that these findings are based on the analysis of transcription factors, their target genes, and the interactions between them, which form the basis of gene regulatory networks. Further investigations may be required to elucidate the specific factors contributing to the observed differences in GRN similarities across the liver lobule and their potential implications for liver function and disease pathogenesis. This spatial variation in transcription factor-target gene interactions reflects the functional specialization of hepatocytes along the central-portal axis, with cells in the central lobule displaying more interactions compared to those in the portal lobule. These findings underscore the intricate regulation of gene expression within the liver lobule and provide insights into the spatial organization of transcriptional networks governing hepatic physiology and response to environmental perturbations.

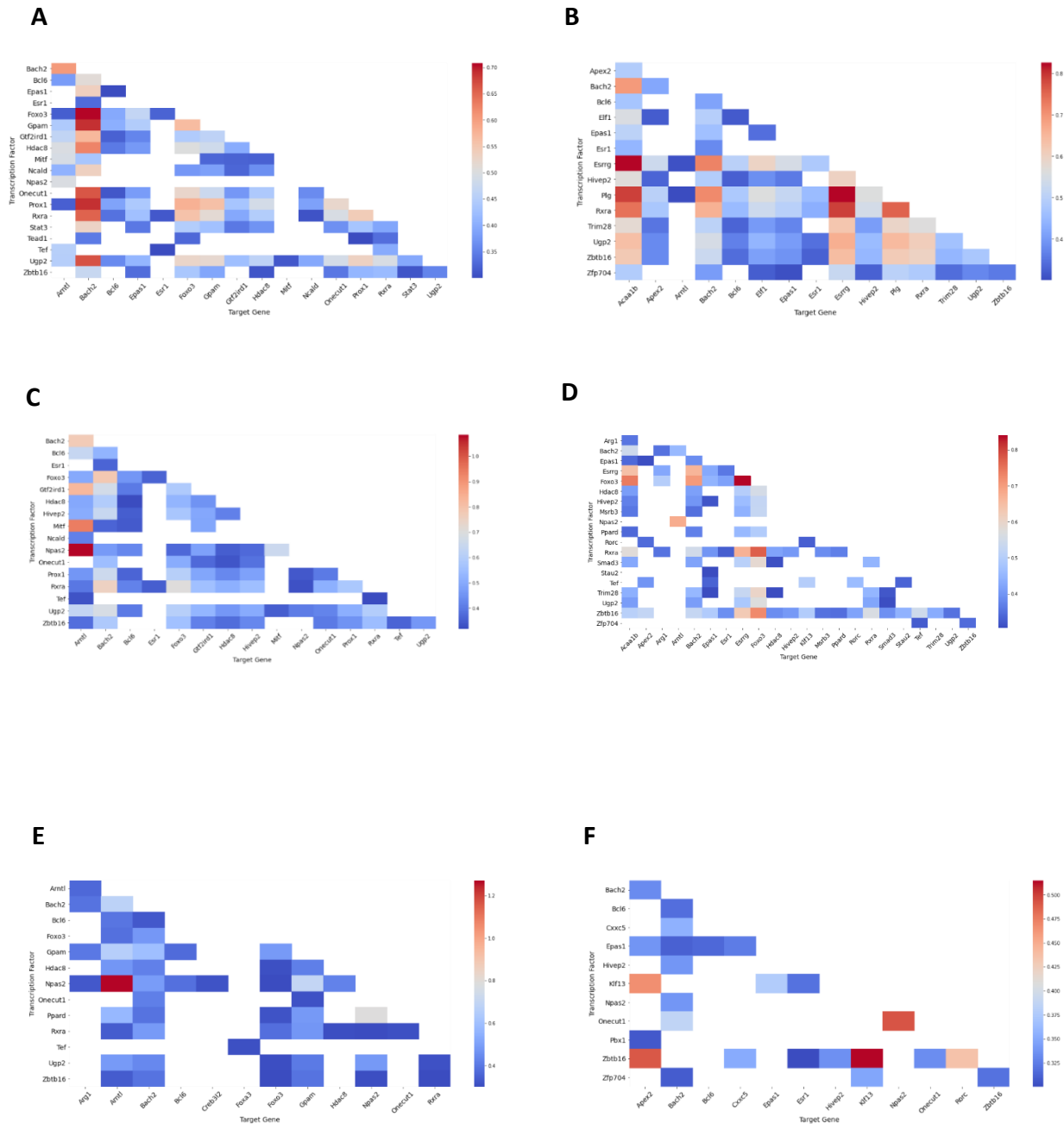
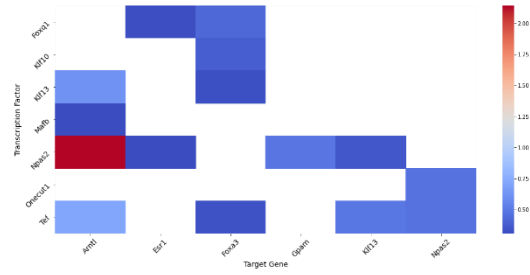


Figure 36: Weighted gene regulatory network in the form of a heatmap to understand the interactions between transcription factors and their weight of interaction for (A) Control Cells in Layer 1 (central axis), (B) Treated cells in Layer 1 (central axis), (C) Control Cells in Layer 2, (D) Treated cells in Layer 2, (E) Control Cells in Layer 3, (F) Treated cells in Layer 3, (G) Control Cells in Layer 4, (H) Treated cells in Layer 4, (I) Control Cells in Layer 5 (portal axis) and (J) Treated cells in Layer 5 (portal axis).

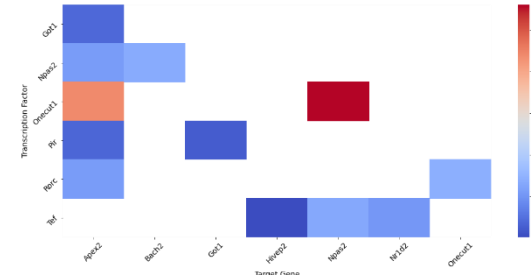
Figure 36 (cont'd)

(K) Average similarity interactions of gene regulatory inference networks between layers of cells in the liver lobule. Z1Z2 refers to average similarities between Zone 1 (layer 1) and Zone 2 (layer 2). Blue squares represent control average interactions (similarities) and red pentagon represent treated average interactions (similarities).

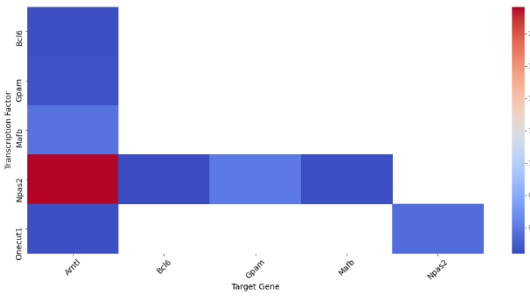
G



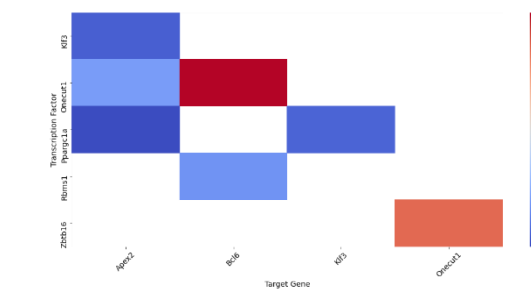
H



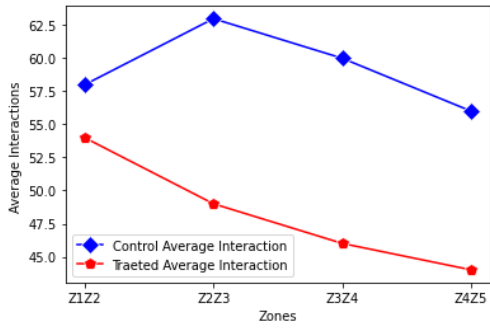
I



J



K



DISCUSSION AND CONCLUSION

Transcription factor (TF)-DNA binding patterns play a crucial role in regulating gene expression¹⁴⁶. However, the precise DNA-binding sequences and the degree of flexibility in these sequences remain elusive for many TFs, including BMAL1, a key regulator of the circadian clock. While the core clock gene regulatory network in the brain's suprachiasmatic nucleus is believed to be similar to that in peripheral tissues, clock-controlled gene expression exhibits significant tissue specificity^{46,147}. In this study, I aimed to elucidate the determinants of BMAL1-DNA binding beyond the simple DNA binding site sequence, leveraging features such as DNA shape and histone modifications. I employed used XGBoost, an ensemble decision tree-based machine learning algorithm, to predict the binding of BMAL1 to its putative binding motif (the E-box) in three mouse tissues – liver, heart and kidney. We developed three different types of models: 1) sequence-only, 2) sequence plus DNA shape, and 3) sequence plus DNA shape plus histone modifications (Fig 2A-B). Examining of the binding motifs revealed that the canonical 5'-CACGTG-3' E-box type was the least frequent among all E-box types in the entire mouse genome and in accessible chromatin regions across the liver, heart, and kidney tissues. However, this E-box type exhibited the highest frequency of BMAL1 binding (Fig 1E), consistent with previous observations that CACGTG is the preferred binding motif for BMAL1¹⁴⁸. Despite the over-representation of the CG dinucleotide at the center of BMAL1-bound E-boxes, these nucleotides did not enhance model performance. Interestingly, the heart tissue harbored more E-boxes in accessible chromatin regions than the liver and kidney, yet it had the lowest number of bound E-boxes in accessible chromatin. The role of circadian rhythms in cardiac function is not well understood, and only 6% of protein-coding genes in the mouse heart exhibit circadian regulation, compared to 11%-16% in the liver¹⁴⁹. This observation is likely a consequence of the overall lower BMAL1 binding in the heart tissue.

However, the underlying reasons for the low level of BMAL1 binding to otherwise accessible E-boxes in the heart remain unexplained. One possible explanation for the low BMAL1 binding to accessible E-boxes in the heart tissue could be the interference of a heart-specific E-box binding factor. For instance, it has been shown that elevated levels of Usf1, a ubiquitous transcription factor, can impede the binding of a mutant CLOCK Δ 19:BMAL1 complex to E-box sites, and other such factors may exist¹⁵⁰. Interestingly, neither the kidney nor the heart within-tissue models achieved the same level of performance as the liver within-tissue model. This could be due to evidence suggesting that the heart circadian rhythm might be phase-shifted compared to the liver, indicating that maximal BMAL1 binding in the heart might occur at a different time than the Zeitgeber time 6 (ZT06)⁴⁶ used for the liver BMAL1 ChIP-seq experiment. Consequently, some of the heart BMAL1-bound E-boxes might have been labeled as unbound, affecting model learning and resulting in lower performance. A limitation of this work is that the authors considered only E-boxes in accessible chromatin regions and disregarded inaccessible E-boxes. While their observations confirmed that, on average, more than 75% of BMAL1 peaks lie in accessible chromatin, suggesting a higher likelihood of BMAL1 binding in these regions, it has been demonstrated that the BMAL1-CLOCK complex can act as a pioneering factor and rhythmically control the accessibility of chromatin surrounding BMAL1-bound sites¹⁵¹. Recent studies have demonstrated that incorporating DNA shape features computed from core transcription factor (TF) binding motifs and their flanking sequences improves the prediction of TF binding for many human TFs^{50,62,63}. Furthermore, DNA topology is highly correlated with the structure and stability of the nucleosome, suggesting that topological changes can influence the binding of TFs to DNA¹⁵². In the sequence plus shape models developed in this study, the DNA shape features EP, Roll, MGW, and ProT exhibited the highest influence on the prediction of bound E-boxes. A recent

study⁵⁰ showed that for Max, a basic helix-loop-helix (bHLH) TF like BMAL1 and CLOCK, it was Roll and ProT that were the dominant determinants of TF-DNA binding affinity. These observations agree with our findings. Further, in agreement with previous studies, I found that DNA shape features by themselves do not enhance model accuracy. The analysis of feature importance in the sequence plus shape plus histone modifications (HMs) models revealed that the HMs H3K27ac and H3K4me3, along with the DNA shape feature EP, dominated binding prediction across the kidney and heart tissues, and were also highly ranked in the liver. Previous studies have shown that H3 acetylation and methylation modifications surrounding CLOCK-BMAL1 bound sites exhibit rhythmic changes¹⁵³. My results findings indicate that even with a single snapshot of these HMs from unsynchronized mice, we can accurately distinguish between bound and unbound E-boxes, likely owing to the information encoded in the shape and surrounding sequence of the E-box motif, in addition to the average levels of histone modifications. I propose that this is likely due to the information that is encoded in the shape and flanking sequence of the E-box motif in addition to the average levels of histone acetylation and methylation. The analysis suggest that this information is tissue-specific, as evidenced by the performance of the cross-tissue models. Intriguingly, the DNA sequence features alone had little to no effect on binding prediction in the kidney and heart. However, the second nucleotide upstream of the E-box had a significant contribution to predicting BMAL1-DNA binding in the liver, with the nucleotide G at this position contributing to approximately 50% of the feature importance score. Analysis of the bound E-box motifs and their flanking sequences revealed an enrichment of the G nucleotide at the third position of the 5' flanking sequence in the liver. Since the heart and kidney models do not rely on this feature, it is understandable that the liver_kidney and liver_heart cross-tissue models showed an unexpected decrease in performance when DNA shape and histone modification features were

added to the sequence features. On the other hand, the kidney_liver and heart_liver cross-tissue models exhibited a boost in performance with the addition of histone modification features. These results suggest that there is some degree of commonality in BMAL1 binding between different tissues. However, in cross-tissue models, sequence-only models exhibited the most robust performance, with the exception of the kidney_heart and heart_kidney models. This indicates that DNA shape and chromatin context features can exhibit a high degree of tissue specificity and are more similar between the kidney and heart than between the liver and the other two tissues. Previous studies on the yeast bHLH transcription factors Cbfl and Tye7 have shown that E-box binding specificity is governed by sequences flanking the E-box, as reflected in DNA shape¹⁵⁴. My findings extend this concept, indicating that not only might DNA shape and chromatin context confer different binding specificities to different transcription factors within the same tissue, but they might also confer different binding specificities to the same transcription factor across different tissues. These observations suggest that while there may be some shared determinants of BMAL1 binding across tissues, there are also tissue-specific factors, such as DNA shape and chromatin context, that modulate BMAL1's binding specificity. This tissue-specific regulation could contribute to the observed differences in clock-controlled gene expression patterns among various tissues.

Dynamic modeling of biological processes from gene regulatory to multicellular network levels provides critical insights into the fundamental properties, physiology, and behaviors related to circadian rhythmicity¹⁵⁵. While experimental and theoretical explorations have extensively detailed the circadian clock gene regulatory network¹⁵⁶, few studies have examined the cell-cell communication processes enabling synchronization of circadian period, amplitude, and phase between autonomous cellular oscillators. Elucidating circadian synchronization through cell-cell

coupling can advance our understanding of circadian rhythm robustness and plasticity at the tissue level²⁹. Neurotransmitters, acting as coupling factors, have been shown to regulate the synchronization mechanism in the suprachiasmatic nucleus (SCN)^{105,157}. However, the coupling factors in other peripheral tissues are unknown. In this study, I developed mathematical models of the mouse hepatic circadian clock to examine intercellular communication and synchronization of autonomous oscillators across the murine liver lobule. The models incorporated core clock genes and their regulatory interactions in individual hepatocytes. Simulations showed that incorporating cell-cell coupling led to synchronized gene expression between hepatocytes, matching experimental findings^{29,158}. Strong synchronicity of circadian oscillation has been associated with period lengthening. However, the models suggest that without synchronization, the period is variable outside of the near 24-hr range. Therefore, optimal cell-cell coupling is required to achieve both synchronicity between cells and appropriate oscillatory periods¹⁵⁹. Synchronicity of circadian oscillation has been shown to induce key cell-cycle events, including cyclin-dependent kinase network activation, cell growth, DNA replication, and cytokinesis¹⁶⁰. A weak synchronicity at the cellular level in the SCN manifests as mistimed sleep and impaired cognitive and psychomotor performance in humans. This circadian misalignment has detrimental effects on physiology and behavior, including deficits in reaction time, memory, alertness, and mood⁸. The sensitivity analysis revealed dependencies between clock components; for instance, increased *Per* transcription decreased *Cry* expression, likely due to their mutual repression¹⁶¹. Specifically, increased *Cry* transcription rates markedly diminished *Per* mRNA levels. However, perturbations in the *Per* transcription rate did not comparably suppress Cryptochrome transcripts³⁷. This aligns with experimental reporter assays demonstrating that the repression strength of PER proteins on Clock/Bmal1-driven transcription is weaker relative to CRY¹⁶². The disproportionate parametric

sensitivities suggest that CRY dynamics play a more dominant role than PER in governing circadian rhythmicity via interlocking negative feedback loops. Validating the coupled mechanism in the model, the authors used bifurcation diagrams to show the periodic stability of the transcription rate parameters. These theoretical bifurcation diagrams generate experimentally testable hypotheses, potentially through overexpression or knockdown of circadian factors in vivo and in vitro, to validate predicted changes in periodicity, amplitude, or phase shifts¹⁴⁹. Our knowledge of the complex mammalian circadian clock mechanism is still incomplete. More than 40 genes directly interact with the core clock genes in generating the circadian oscillatory rhythm¹⁶³. It is, therefore, essential to understand the temporal and spatial dynamics and the regulatory mechanism of the circadian clock oscillation. I addressed this knowledge gap with a detailed mathematical model incorporating known clock and associated rhythmic genes. The current model focused on healthy cells under normal circadian entrainment. An important extension would be to model circadian disruption by genetic alterations or toxicant exposure. The model could be used to predict the effects of parameter changes representing mutations or cellular damage. Linking the circadian clock model with zoned metabolism models could offer insights into compounding hepatic effects across scales. While the study provides valuable insights into the circadian clock mechanism, I acknowledge the limitations of my current model, which focuses on healthy cells under normal circadian entrainment. I propose extending the model to simulate circadian disruption caused by genetic alterations or exposure to toxicants, which could be achieved by modifying parameters to represent mutations or cellular damage. Additionally, integrating the circadian clock model with zoned metabolism models could provide a holistic understanding of the multi-scale effects on hepatic function.

The liver exhibits a remarkably intricate spatiotemporal metabolic organization. As the primary site of drug metabolism, elucidating the impact of chemical compounds on hepatic gene regulation in both spatial and temporal domains could provide critical insights. Firstly, characterizing the acute restructuring of rhythmic and zonal gene expression patterns in response to drug exposure could reveal adaptive mechanisms. Secondly, disruption of these regulatory mechanisms likely contributes to the progression of pathological states such as non-alcoholic fatty liver disease. Analysis of drug effects on the liver has often been limited by examining temporal rhythmicity and spatial zonation independently. In this study, we integrated these hepatic properties to account for chemical perturbations, an important consideration given the understudied aspect of how the liver reorganizes its metabolic framework in response to toxic stimuli. We expanded on previous work unifying rhythmicity and zonation to incorporate the effects of acute exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD). Despite the short timescale, we demonstrated that TCDD significantly impacts both rhythms and zonation. The canonical TCDD receptor, the aryl hydrocarbon receptor (AhR), exhibited both rhythmic and zonal expression patterns. Moreover, numerous TCDD-induced effects had both temporal and spatial components. Overall, this integrated methodology enabled a more comprehensive characterization of rapid liver reorganization in response to TCDD exposure than studying zonation or rhythms in isolation. Our findings highlight the utility of integrating diverse regulatory properties when elucidating the impacts of chemical compounds on hepatic gene regulation and metabolic organization. Our integrated analysis demonstrates that hepatic rhythmicity is more sensitive to acute exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) compared to zonation. This aligns with previous findings showing that subchronic TCDD administration at the same dose greatly dampened amplitude and significantly phase-shifted the core circadian clock²⁸. Although acute TCDD did

not perturb all core clock genes analyzed, we still observed significant influences on nearly half of these critical circadian regulators. Effects were most notable for genes directly downstream of the CLOCK-ARNTL heterodimer (e.g., *Per2*, *Nr1d2*), supporting the hypothesis that the aryl hydrocarbon receptor (AhR) interferes with CLOCK-ARNTL transcriptional activation by competing for ARNTL binding. Additional evidence lies in the expression pattern of *Per1*, a direct target inhibited by AHR-ARNTL binding. Although *Per1*, a core circadian gene, was excluded from our dataset due to low variance, its normalized expression showed signs of repression at multiple timepoints between 2-24 hours post-TCDD exposure, consistent with disrupted CLOCK-ARNTL control. TCDD elicited variable effects on rhythmic hepatic genes; while most exhibited modest expression changes at specific times, overall oscillation patterns were largely maintained with treatment for approximately 85% of rhythmic transcripts. However, approximately 15% of rhythmic genes gained or lost transcriptional rhythmicity, representing selective arrhythmic effects. Notably, gene set enrichment analysis linked these arrhythmia-gaining genes to canonical TCDD response pathways, including hallmark AhR-mediated responses^{126,164}. This indicates that TCDD preferentially disrupts rhythmicity of major regulatory nodes and downstream processes most susceptible to aberrant AhR activation. Rather than inducing widespread distortions in cyclic waveforms, TCDD appears to ablate oscillatory control of TCDD-vulnerable pathways. TCDD also significantly impacted zonal gene regulation, altering spatial expression patterns for nearly 25% of zonated transcripts. Notably, about 30% of these genes either gained or lost zonation, representing substantial chemical-induced re-patterning. Genes losing zonation were enriched for phase II UDP-glucuronosyltransferase metabolism and localized to the central hepatic layer, matching the enrichment of the AhR receptor itself. TCDD-perturbed genes were also enriched for known regulators of zonation, potentially via crosstalk with Wnt/ β -catenin signaling. This was

most prominent for genes jointly exhibiting rhythmic and zonal properties, which are hypothesized to arise from Wnt-mediated pericentral signals. As AhR shows central layer enrichment, its aberrant activation by TCDD could disrupt this external cue, explaining the observed perturbation in rhythmicity and zonation. While the precise mechanisms through which aberrant AhR signaling perturbs hepatic rhythmicity and zonation after TCDD exposure remain unclear, potential routes could involve direct transcriptional regulation via AhR binding to target gene cis-regulatory elements or hierarchical effects by disrupting core clock regulators, secondarily altering cyclic control of downstream processes. High-throughput approaches like ChIP-seq, profiling genome-wide binding locations for AhR, ARNTL, and binding partners after TCDD exposure, could elucidate these questions and map binding sites to clarify pathways exhibiting direct, canonical AhR-mediated regulation versus downstream, non-canonical disruption. In summary, these results indicate that the disruption of temporal rhythmicity is a primary route through which TCDD alters hepatic gene regulation, relative to spatial patterning. The specific effects on core circadian components highlight direct mechanisms through which aberrant AhR activation propagates to disrupt hepatic rhythmicity, with the immediate downstream targets of CLOCK-ARNTL appearing especially sensitive to functional interference by ligand-activated AhR. This approach enables matching of precise TCDD-induced expression changes to interactions within the circadian regulatory network, refining hypotheses on the routes by which AhR signaling disrupts clock function. The method enables comparative assessment of the impacts of chemical exposure on rhythmic versus zonal gene expression, revealing that perturbations in rhythmic transcriptional outputs precede and potentially contribute to changes in zonation in the case of TCDD exposure. The developed methodology provides a framework readily extensible to additional chemical exposure contexts, dose-response relationships, and single-nuclei RNA sequencing datasets

assessing hepatic chemical impacts. By matching pathway-specific expression changes to overarching spatiotemporal regulatory logic, this systems-based strategy has potential utility for pharmacology and toxicology, enabling researchers to situate perturbed pathways within the broader scheme of metabolic zonation and rhythmicity. Characterizing how xenobiotic disruption of specific functional modules scales up to influence zonation control pathways and circadian timing at the tissue level will drive a more holistic understanding of chemical threats to hepatic function. Overall, this multi-parametric modeling approach provides a blueprint for deep phenotyping of gene regulatory restructuring in this vital metabolic organ.

BIBLIOGRAPHY

1. Huang, T. C., Tu, J., Chow, T. J. & Chen, T. H. Circadian Rhythm of the Prokaryote *Synechococcus* sp. RF-1. *Plant Physiol* **92**, 531 (1990).
2. Hastings, M. H. *Circadian Clocks. Current biology : CB* vol. 7 (1997).
3. Ko, C. H. & Takahashi, J. S. Molecular components of the mammalian circadian clock. *Hum Mol Genet* **15 Spec No 2**, (2006).
4. Observation botanique. Histoire de l'Academie Royale de Sciences (Paris). https://www.researchgate.net/publication/291798032_Observation_botanique_Histoire_de_l'Academie_Royale_de_Sciences_Paris.
5. A Behavioristic Study of the Activity of the Rat. <https://psycnet.apa.org/record/1926-07705-001>.
6. Endogenous Rhythms in Insects and Microorganisms on JSTOR. <https://www.jstor.org/stable/2458417>.
7. Vitaterna, M. H., Shimomura, K. & Jiang, P. Genetics of Circadian Rhythms. *Neurol Clin* **37**, 487–504 (2019).
8. Patton, A. P. & Hastings, M. H. The Mammalian Circadian Time-Keeping System. *J Huntingtons Dis* **12**, 91 (2023).
9. Gonze, D. Modeling circadian clocks: Roles, advantages, and limitations. *Cent Eur J Biol* **6**, 712–729 (2011).
10. Takahashi, J. S. Transcriptional architecture of the mammalian circadian clock. *Nat Rev Genet* **18**, 164–179 (2017).
11. Lee, Y. Roles of circadian clocks in cancer pathogenesis and treatment. *Experimental and Molecular Medicine* vol. 53 1529–1538 Preprint at <https://doi.org/10.1038/s12276-021-00681-0> (2021).
12. Maywood, E. S., O'Neill, J. S., Chesham, J. E. & Hastings, M. H. Minireview: The circadian clockwork of the suprachiasmatic nuclei--analysis of a cellular oscillator that drives endocrine rhythms. *Endocrinology* **148**, 5624–5634 (2007).
13. Skene, D. J. & Arendt, J. *Human Circadian Rhythms: Physiological and Therapeutic Relevance of Light and Melatonin. Ann Clin Biochem* vol. 43 (2006).
14. Schibler, U. *et al.* Clock-Talk: Interactions between central and peripheral circadian oscillators in mammals. *Cold Spring Harb Symp Quant Biol* **80**, 223–232 (2016).
15. Mohawk, J. A., Green, C. B. & Takahashi, J. S. Central and peripheral circadian clocks in mammals. *Annu Rev Neurosci* **35**, 445–462 (2012).
16. Albrecht, U. Timing to perfection: the biology of central and peripheral circadian clocks. *Neuron* **74**, 246–260 (2012).

17. Gamble, K. L., Berry, R., Frank, S. J. & Young, M. E. Circadian clock control of endocrine factors. *Nat Rev Endocrinol* **10**, 466–475 (2014).
18. Green, C. B., Takahashi, J. S. & Bass, J. The meter of metabolism. *Cell* **134**, 728–742 (2008).
19. Richards, J. & Gumz, M. L. Mechanism of the circadian clock in physiology. *Am J Physiol Regul Integr Comp Physiol* **304**, 1053–1064 (2013).
20. Esquirol, Y. *et al.* Shift work and metabolic syndrome: Respective impacts of job strain, physical activity, and dietary rhythms. *Chronobiol Int* **26**, 544–559 (2009).
21. Biggi, N., Consonni, D., Galluzzo, V., Sogliani, M. & Costa, G. Metabolic syndrome in permanent night workers. *Chronobiol Int* **25**, 443–454 (2008).
22. Hadadi, E. *et al.* Chronic circadian disruption modulates breast cancer stemness and immune microenvironment to drive metastasis in mice. *Nat Commun* **11**, (2020).
23. Masri, S. & Sassone-Corsi, P. The emerging link between cancer, metabolism, and circadian rhythms. *Nat Med* **24**, 1795–1803 (2018).
24. Castanon-Cervantes, O. *et al.* Dysregulation of inflammatory responses by chronic circadian disruption. *J Immunol* **185**, 5796–5805 (2010).
25. Maury, E., Ramsey, K. M. & Bass, J. Circadian rhythms and metabolic syndrome: from experimental genetics to human disease. *Circ Res* **106**, 447–462 (2010).
26. Williamson, M. A., Gasiewicz, T. A. & Opanashuk, L. A. Aryl hydrocarbon receptor expression and activity in cerebellar granule neuroblasts: Implications for development and dioxin neurotoxicity. *Toxicological Sciences* **83**, 340–348 (2005).
27. Mimura, J. & Fujii-Kuriyama, Y. Functional role of AhR in the expression of toxic effects by TCDD. *Biochim Biophys Acta Gen Subj* **1619**, 263–268 (2003).
28. 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**, 1–18 (2019).
29. Droin, C. *et al.* Space-time logic of liver gene expression at sub-lobular scale. *Nat Metab* **3**, 43–58 (2021).
30. Wang, Y. *et al.* A proteomics landscape of circadian clock in mouse liver. *Nature Communications* **2018 9:1** **9**, 1–16 (2018).
31. DeBruyne, J. P., Weaver, D. R. & Dallmann, R. The hepatic circadian clock modulates xenobiotic metabolism in mice. *J Biol Rhythms* **29**, 277 (2014).
32. Ferrell, J. M. & Chiang, J. Y. L. Circadian rhythms in liver metabolism and disease. *Acta Pharm Sin B* **5**, 113 (2015).
33. Annunziato, S. & Tchorz, J. S. Liver zonation—a journey through space and time. *Nat Metab* **3**, 7–8 (2021).

34. Jungermann, K. Zonation of metabolism and gene expression in liver. *Histochem Cell Biol* **103**, 81–91 (1995).
35. Burke, Z. D. & Tosh, D. The Wnt/ β -catenin pathway: master regulator of liver zonation? *BioEssays* **28**, 1072–1077 (2006).
36. Mukherji, A., Bailey, S. M., Staels, B. & Baumert, T. F. The circadian clock and liver function in health and disease. *J Hepatol* **71**, 200–211 (2019).
37. Relógio, A. *et al.* Tuning the mammalian circadian clock: Robust synergy of two loopsRelógio, A., Westermarck, P. O., Wallach, T., Schellenberg, K., Kramer, A., & Herzel, H. (2011). Tuning the mammalian circadian clock: Robust synergy of two loops. *PLoS Computational Biology*,. *PLoS Comput Biol* **7**, 1–18 (2011).
38. Forger, D. B. & Peskin, C. S. A detailed predictive model of the mammalian circadian clock. *Proc Natl Acad Sci U S A* **100**, 14806–14811 (2003).
39. Takahashi, J. S., Hong, H. K., Ko, C. H. & McDearmon, E. L. The genetics of mammalian circadian order and disorder: implications for physiology and disease. *Nature Reviews Genetics* **2008 9:10 9**, 764–775 (2008).
40. Landgraf, D., Wang, L. L., Diemer, T. & Welsh, D. K. NPAS2 Compensates for Loss of CLOCK in Peripheral Circadian Oscillators. *PLoS Genet* **12**, e1005882 (2016).
41. Cox, K. H. & Takahashi, J. S. Circadian clock genes and the transcriptional architecture of the clock mechanism. *J Mol Endocrinol* **63**, R93–R102 (2019).
42. Yoo, S. H. *et al.* A noncanonical E-box enhancer drives mouse Period2 circadian oscillations in vivo. *Proc Natl Acad Sci U S A* **102**, 2608–2613 (2005).
43. Kathiresan, S. & Srivastava, D. Genetics of Human Cardiovascular Disease. *Cell* **148**, 1242 (2012).
44. Schödel, J. *et al.* Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. *Nat Genet* **44**, 420–425 (2012).
45. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
46. Beytebiere, J. R. *et al.* Tissue-specific BMAL1 cistromes reveal that rhythmic transcription is associated with rhythmic enhancer–enhancer interactions. *Genes Dev* **33**, 294–309 (2019).
47. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res* **25**, 1268–1280 (2015).
48. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol* **47**, 1–8 (2017).
49. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22**, 1798–1812 (2012).

50. Zhou, T. *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A* **112**, 4654–4659 (2015).
51. Filipovic, D. *et al.* Interpretable predictive models of genome-wide aryl hydrocarbon receptor-DNA binding reveal tissue specific binding determinants. *Toxicol Sci* (2023) doi:10.1093/TOXSCI/KFAD094.
52. Steuernagel, L. *et al.* Computational identification of tissue-specific transcription factor cooperation in ten cattle tissues. *PLoS One* **14**, e0216475 (2019).
53. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
54. Arvey, A., Agius, P., Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* **22**, 1723–1734 (2012).
55. Gordân, R. *et al.* Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Rep* **3**, 1093–1104 (2013).
56. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21**, 447–455 (2011).
57. Das, M. K. & Dai, H. K. A survey of DNA motif finding algorithms. *BMC Bioinformatics* **8**, 1–13 (2007).
58. Alipanahi, B., DeLong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **2015 33:8** **33**, 831–838 (2015).
59. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* **44**, (2016).
60. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* **2015 12:10** **12**, 931–934 (2015).
61. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-August-2016**, 785–794 (2016).
62. Mathelier, A. *et al.* DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst* **3**, 278-286.e4 (2016).
63. Wang, S. *et al.* Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture. *Mol Ther Nucleic Acids* **24**, 154–163 (2021).
64. Wang, Y., Li, X. & Hu, H. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics* **103**, 222–228 (2014).
65. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* **39**, 381–399 (2014).

66. Li, J. *et al.* Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res* **45**, 12877 (2017).
67. Benveniste, D., Sonntag, H. J., Sanguinetti, G. & Sproul, D. Transcription factor binding predicts histone modifications in human cell lines. *Proc Natl Acad Sci U S A* **111**, 13367–13372 (2014).
68. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
69. Chiu, T. P. *et al.* DNASHapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* **32**, 1211–1213 (2016).
70. Zhou, T. *et al.* DNASHape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* **41**, (2013).
71. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160–W165 (2016).
72. Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618–1619 (2014).
73. Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. An Introduction to Logistic Regression Analysis and Reporting. <https://doi.org/10.1080/00220670209598786> **96**, 3–14 (2010).
74. ENCODE. <https://www.encodeproject.org/>.
75. Marri, D., Filipovic, D., Kana, O., Tischkau, S. & Bhattacharya, S. Prediction of mammalian tissue-specific CLOCK–BMAL1 binding to E-box DNA motifs. *Sci Rep* **13**, (2023).
76. Liu, S. *et al.* Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility. *BMC Bioinformatics* **18**, 1–11 (2017).
77. Benveniste, D., Sonntag, H. J., Sanguinetti, G. & Sproul, D. Transcription factor binding predicts histone modifications in human cell lines. *Proc Natl Acad Sci U S A* **111**, 13367–13372 (2014).
78. Guccione, E. *et al.* Myc-binding-site recognition in the human genome is determined by chromatin context. *Nature Cell Biology* 2006 8:7 **8**, 764–770 (2006).
79. Xin, B. & Rohs, R. Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Res* **28**, 321–333 (2018).
80. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 2009 459:7243 **459**, 108–112 (2009).
81. Ramsey, S. A. *et al.* Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics* **26**, 2071–2075 (2010).
82. Korobov, M. & Lopuhin, K. ELI5 Documentation Release 0.11.0. (2021).
83. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190 (2004).

84. Reppert, S. M. & Weaver, D. R. Coordination of circadian timing in mammals. *Nature* 2002 418:6901 **418**, 935–941 (2002).
85. Moore, R. Y., Speh, J. C. & Leak, R. K. Suprachiasmatic nucleus organization. *Cell Tissue Res* **309**, 89–98 (2002).
86. Maronde, E. *et al.* The clock genes Period 2 and Cryptochrome 2 differentially balance bone formation. *PLoS One* **5**, (2010).
87. Takahashi, J. S. Molecular Neurobiology and Genetics of Circadian Rhythms in Mammals. <https://doi.org/10.1146/annurev.ne.18.030195.002531> **18**, 531–553 (2003).
88. Lamia, K. A. & Evans, R. M. Metabolism: Tick, tock, a beta-cell clock. *Nature* **466**, 571–572 (2010).
89. Sahar, S. & Sassone-Corsi, P. Metabolism and cancer: the circadian clock connection. *Nat Rev Cancer* **9**, 886–896 (2009).
90. Hastings, M. H., Reddy, A. B. & Maywood, E. S. A clockwork web: circadian timing in brain and periphery, in health and disease. *Nature Reviews Neuroscience* 2003 4:8 **4**, 649–661 (2003).
91. Marcheva, B. *et al.* Disruption of the clock components CLOCK and BMAL1 leads to hypoinsulinaemia and diabetes. *Nature* **466**, 627–631 (2010).
92. Hardin, P. E. & Panda, S. Circadian timekeeping and output mechanisms in animals. *Curr Opin Neurobiol* **23**, 724–731 (2013).
93. Guillaumond, F., Dardente, H., Giguère, V. & Cermakian, N. Differential control of Bmal1 circadian transcription by REV-ERB and ROR nuclear receptors. *J Biol Rhythms* **20**, 391–403 (2005).
94. Gebhardt, R. Metabolic zonation of the liver: regulation and implications for liver function. *Pharmacol Ther* **53**, 275–354 (1992).
95. Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. (2017) doi:10.1038/nature21065.
96. Li, Y., Liu, Z., Luo, J. & Wu, H. Coupling-induced synchronization in multicellular circadian oscillators of mammals. *Cogn Neurodyn* **7**, 59–65 (2013).
97. Schmal, C., Herzog, E. D. & Herzog, H. Measuring Relative Coupling Strength in Circadian Systems. *J Biol Rhythms* **33**, 84–98 (2018).
98. Finger, A. M. *et al.* Intercellular coupling between peripheral circadian oscillators by TGF- β signaling. *Sci Adv* **7**, 1–20 (2021).
99. Sloin, H. E. *et al.* Interactions between the circadian clock and TGF- β signaling pathway in zebrafish. *PLoS One* **13**, e0199777 (2018).
100. Kon, N. *et al.* Activation of TGF- β /activin signalling resets the circadian clock through rapid induction of Dec1 transcripts. *Nature Cell Biology* 2008 10:12 **10**, 1463–1469 (2008).

101. Shirakawa, T., Honma, S. & Honma, K. I. Multiple oscillators in the suprachiasmatic nucleus. *Chronobiol Int* **18**, 371–387 (2001).
102. Cholicco Giovanni N. Time-Course Single-Nuclei RNA-Seq Analysis Identifies NRG/ERBB Signaling Dysregulation in TCDD-Induced Hepatotoxicity. (2023).
103. Finger, A. M., Dibner, C. & Kramer, A. Coupled network of the circadian clocks: a driving force of rhythmic physiology. *FEBS Lett* **594**, 2734–2769 (2020).
104. Robles, M. S., Humphrey, S. J., Mann Correspondence, M. & Mann, M. Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology Cell Metabolism Resource Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology. *Cell Metab* **25**, 118–127 (2017).
105. Gonze, D., Bernard, S., Waltermann, C., Kramer, A. & Herzog, H. Spontaneous synchronization of coupled circadian oscillators. *Biophys J* **89**, 120–129 (2005).
106. Brenna, A. & Albrecht, U. Phosphorylation and Circadian Molecular Timing. *Front Physiol* **11**, 612510 (2020).
107. Lei, J. Mathematical Models for Gene Regulatory Network Dynamics. 145–198 (2021) doi:10.1007/978-3-030-73033-8_5.
108. Swat, M. H. *et al.* Multi-Scale Modeling of Tissues Using CompuCell3D. *Methods Cell Biol* **110**, 325–366 (2012).
109. Swat, M. H. *et al.* *Modeling of Tissues Using CompuCell3D Methods*. (2012). doi:10.1016/B978-0-12-388403-9.00013-8.Multi-Scale.
110. Appendix — CC3D Reference Manual 4.2.4 documentation. <https://compucell3dreferencemanual.readthedocs.io/en/latest/appendix.html>.
111. Kim, R. & Reed, M. C. A mathematical model of circadian rhythms and dopamine. *Theor Biol Med Model* **18**, (2021).
112. Finger, A. Molecular Mechanisms of Intercellular Coupling among Peripheral Circadian Oscillators. (2020).
113. Balsa-Canto, E. & Banga, J. R. AMIGO: A model identification toolbox based on global optimization and its applications in biosystems. in *IFAC Proceedings Volumes (IFAC-PapersOnline)* vol. 11 132–137 (IFAC Secretariat, 2010).
114. Balsa-Canto, E., Henriques, D., Gábor, A. & Banga, J. R. *AMIGO 2 THEORETICAL BACKGROUND*. (2016).
115. 167293.167637.
116. Battogtokh, D. & Tyson, J. J. Deciphering the Dynamics of Interlocked Feedback Loops in a Model of the Mammalian Circadian Clock. *Biophys J* **115**, 2055–2066 (2018).
117. XPP/XPPAUT Homepage. <https://sites.pitt.edu/~phase/bard/bardware/xpp/xpp.html>.
118. Leloup, J. C. & Goldbeter, A. Modeling the mammalian circadian clock: sensitivity analysis and multiplicity of oscillatory mechanisms. *J Theor Biol* **230**, 541–562 (2004).

119. Egea, J. A., Martí, R. & Banga, J. R. An evolutionary method for complex-process optimization. *Comput Oper Res* **37**, 315–324 (2010).
120. Villaverde, A. F. *et al.* BioPreDyn-bench: A suite of benchmark problems for dynamic modelling in systems biology. *BMC Syst Biol* **9**, (2015).
121. 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).
122. Santostefano, M. J. *et al.* Dose-Dependent Localization of TCDD in Isolated Centrilobular and Periportal Hepatocytes. (1999).
123. Wild, S. L. *et al.* The Canonical Wnt Pathway as a Key Regulator in Liver Development, Differentiation and Homeostatic Renewal. *Genes (Basel)* **11**, 1–20 (2020).
124. Nault, R. *et al.* Single-cell transcriptomics shows dose-dependent disruption of hepatic zonation by TCDD in mice. *Toxicol Sci* **191**, 135–148 (2023).
125. Huang, P., Tofighi, R., Emgard, M. & Ceccatelli, S. Cell death induced by 2,3,7,8-tetrachlorodibenzo-p-dioxin (2,3,7,8-TCDD) in AtT-20 pituitary cells. *Toxicology* **207**, 391–399 (2005).
126. 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. *Cell Mol Gastroenterol Hepatol* **11**, 147–159 (2021).
127. Cholico, G. N. *et al.* Consequences of reprogramming acetyl-CoA metabolism by 2,3,7,8-tetrachlorodibenzo-p-dioxin in the mouse liver. *Scientific Reports 2023 13:1* **13**, 1–14 (2023).
128. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, (2018).
129. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods 2018 15:12* **15**, 1053–1058 (2018).
130. Nelder, J. A. & Mead, R. A simplex method for function minimization. *Comput J* **7**, 308–313 (1965).
131. Schwarz, G. Estimating the Dimension of a Model. <https://doi.org/10.1214/aos/1176344136> **6**, 461–464 (1978).
132. Thaben, P. F. & Westermark, P. O. Differential rhythmicity: detecting altered rhythmicity in biological data. *Bioinformatics* **32**, 2800–2808 (2016).
133. 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* **6**, 750–764 (2022).
134. Coifman, R. R. & Lafon, S. Diffusion maps. *Appl Comput Harmon Anal* **21**, 5–30 (2006).

135. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* 2016 13:10 **13**, 845–848 (2016).
136. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* **17**, (2016).
137. GitHub - naef-lab/dryR: Differential Rhythmicity analysis in R is an R package that provides the statistical framework to assess differential rhythmicity of a time series with two and more conditions. <https://github.com/naef-lab/dryR>.
138. 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).
139. Gougelet, A. *et al.* T-cell factor 4 and β -catenin chromatin occupancies pattern zonal liver metabolism in mice. *Hepatology* **59**, 2344–2357 (2014).
140. Lamb, C. L. *et al.* Aryl Hydrocarbon Receptor Activation by TCDD Modulates Expression of Extracellular Matrix Remodeling Genes during Experimental Liver Fibrosis. *Biomed Res Int* **2016**, (2016).
141. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314–2321 (2017).
142. Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* **24**, 739–754 (2023).
143. Zhang, Q. *et al.* Embracing Systems Toxicology at Single-Cell Resolution. *Curr Opin Toxicol* **16**, 49–57 (2019).
144. Zhang, S., Xiangtao, L. I., Jiecong, L. I. N., Qiuzhen, L. I. N. & Wong, K. C. Review of single-cell RNA-seq data clustering for cell-type identification and characterization. *RNA* **29**, (2023).
145. Keyl, P. *et al.* Single-cell gene regulatory network prediction by explainable AI. *Nucleic Acids Res* **51**, E20–E20 (2023).
146. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* 2003 424:6945 **424**, 147–151 (2003).
147. Mure, L. S. *et al.* Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science (1979)* **359**, (2018).
148. Hogenesch, J. B., Gu, Y. Z., Jain, S. & Bradfield, C. A. The basic-helix-loop-helix-PAS orphan MOP3 forms transcriptionally active complexes with circadian and hypoxia factors. *Proc Natl Acad Sci U S A* **95**, 5474–5479 (1998).
149. Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E. & Hogenesch, J. B. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proc Natl Acad Sci U S A* **111**, 16219–16224 (2014).
150. Shimomura, K. *et al.* Usf1, a suppressor of the circadian Clock mutant, reveals the nature of the DNA-binding of the CLOCK:BMAL1 complex in mice. *Elife* **2**, 426 (2013).

151. Menet, J. S., Pescatore, S. & Rosbash, M. CLOCK:BMAL1 is a pioneer-like transcription factor. *Genes Dev* **28**, 8–13 (2014).
152. Gupta, P., Zlatanova, J. & Tomschik, M. Nucleosome Assembly Depends on the Torsion in the DNA Molecule: A Magnetic Tweezers Study. *Biophys J* **97**, 3150 (2009).
153. Koike, N. *et al.* Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. *Science* **338**, 349 (2012).
154. Grove, C. A. *et al.* A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138**, 314–327 (2009).
155. DeWoskin, D. *et al.* Distinct roles for GABA across multiple timescales in mammalian circadian timekeeping. *Proc Natl Acad Sci U S A* **112**, E3911–E3919 (2015).
156. Agostino, P. V. *et al.* Sildenafil accelerates reentrainment of circadian rhythms after advancing light schedules. *PNAS* **104**, 9834–9839 (2007).
157. Bernard, S., Gonze, D., Čajavec, B., Herzel, H. & Kramer, A. Synchronization-induced rhythmicity of circadian oscillators in the suprachiasmatic nucleus. *PLoS Comput Biol* **3**, 667–679 (2007).
158. Gagliano, O. *et al.* Synchronization between peripheral circadian clock and feeding-fasting cycles in microfluidic device sustains oscillatory pattern of transcriptome. *Nature Communications* 2021 12:1 **12**, 1–12 (2021).
159. Garcia-Ojalvo, J., Elowitz, M. B. & Strogatz, S. H. *Modeling a Synthetic Multicellular Clock: Repressilators Coupled by Quorum Sensing*. www.pnas.org/cgi/doi/10.1073/pnas.0307095101 (2004).
160. Feillet, C. *et al.* Phase locking and multiple oscillating attractors for the coupled mammalian clock and cell cycle. *Proc Natl Acad Sci U S A* **111**, 9828–9833 (2014).
161. Kim, J. K. & Forger, D. B. A mechanism for robust circadian timekeeping via stoichiometric balance. *Mol Syst Biol* **8**, 1–14 (2012).
162. Ye, R. *et al.* Dual modes of CLOCK:BMAL1 inhibition mediated by Cryptochrome and period proteins in the mammalian circadian clock. *Genes Dev* **28**, 1989–1998 (2014).
163. Lehmann, R. *et al.* Assembly of a comprehensive regulatory network for the mammalian circadian clock: A bioinformatics approach. *PLoS One* **10**, (2015).
164. Nault, R. *et al.* Dose-dependent metabolic reprogramming and differential gene expression in TCDD-elicited hepatic fibrosis. *Toxicological Sciences* **154**, 253–266 (2016).