COMPUTATIONAL MODELING OF GENOME-WIDE DNA BINDING AND PROTEIN INTERACTIONS BY THE ARYL HYDROCARBON RECEPTOR

By

David Filipovic

A DISSERTATION

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

Biomedical Engineering – Doctor of Philosophy
Computational Mathematics, Science and Engineering – Dual Major

2023

ABSTRACT

The aryl hydrocarbon receptor (AhR) is a ligand inducible transcription factor (TF) with multiple endogenous and exogenous ligands. AhR regulates many cellular processes including differentiation, development, and xenobiotic metabolism. Among its exogenous ligands 2, 3, 7, 8 tetrachlorodibenzo-p-dioxin (TCDD) is its most potent inducer. Upon ligand binding, inactive cytosolic AhR undergoes a conformational change ultimately leading to its nuclear localization. Within the nucleus, AhR is thought to primarily dimerize with AhR nuclear translocator (ARNT) to form a functional TF which binds to DNA at dioxin response elements (DREs) and regulates transcription of AhR target genes. Most DREs in accessible chromatin are not bound by AhR, and DREs accessible in multiple cell lines or type can be bound in some and unbound in others. Still, since AhR possesses a strong core binding motif 5'-GCGTG-3', it is suited for a motif-centered analysis of its binding. To investigate determinants of AhR binding I developed interpretable machine learning models predicting the binding status of DREs in MCF-7, GM17212, HepG2 cells, and primary human hepatocytes. I conclude that AhR binding is driven by a complex interplay of cell-agnostic DRE flanking sequence and cell-specific local chromatin context.

On the other hand, AhR can bind DNA in absence of ARNT. Both, RelA and KLF6 have been shown to physically interact with AhR and together drive the activation of several genes. For example, the activation of 1) c-myc in breast cancer and 2) PAI-1, p21cip1, and E-cadherin genes is driven by AhR interacting with RelA and KLF6, respectively. However, it is unknown if these interactions with AhR occur genome-wide or if they are localized to a small number of genes. I developed a computational method to investigate protein-protein interactions at AhR-bound sites. Results confirm ARNT as the main dimerization partner of AhR genome-wide in TCDD-exposed MCF-7 cells. By contrast, in untreated HepG2 cells, KLF6 and RelA but not ARNT were the main

dimerization partners of AhR. These findings indicate that the role of AhR is likely ligand-dependent and can potentially be explained through dimerization with different partners.

This thesis is dedicated to all who love me and whom I loved.

ACKNOWLEDGMENTS

I would like to acknowledge and offer my kindest thanks to my advisor and mentor, Dr. Sudin Bhattacharya. I have learned invaluable research and life lessons under the kind tutelage of Dr. Bhattacharya. Thank you for providing me with this amazing opportunity. I appreciate your patience and understanding over the many years we have worked together. I would also like to thank the esteemed members of my committee, Dr. Adam Alessio, Dr. Christopher Contag and Dr. Jianrong Wang. Working with you was such a pleasure and I definitely could not have made it as far as I have without your kind guidance and insightful advice. Dr. Arjun Krishnan, it was an honor having you on my committee. I was sad to see you move, but I am happy to see the next chapter of your life unfold. Thank you all! Additionally, I would like to thank Dr. Rory Connolly, Dr. Almudena Veiga-Lopez, and Dr. Suresh Cuddapah for providing invaluable mentorship and support. Dr. Connolly, thank you for teaching me the ins and outs of pharmacokinetic modeling. Dr. Veiga-Lopez, thank you for teaching me responsibility and accountability, and for providing an amazing role model of a compassionate principal investigator. I have grown immensely for having crossed paths with you and I will be forever grateful. Dr. Cuddapah, thank you for your precise, pinpoint critiques and comments. Even today, I continue to learn from your teachings.

Next, I would like to thank my family and friends for being there with me every step of the way. Special thanks to Evran Ural. Thank you for being there for me! I have gone through some very hard times, but they were all the more easier because you were there to support me. I hope you know how much I appreciate you. Thank you Dr. Wenjie Qi for being an amazing friend and always being there to hear me out. Thank you Dr. Xander Farnum, it was real. Thank you, Dr. Chima Maduka, it was so fun living with you, and I miss talking science with you daily.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

3D – three dimensional

3-MC – methylcholanthrene

ADME – administration, distribution, metabolism, and elimination

AhR – aryl hydrocarbon receptor

AhRE – aryl hydrocarbon response element

AIP – AhR interacting protein

ALDH – aldehyde dehydrogenase

ARNT – aryl hydrocarbon receptor nuclear translocator

auROC – area under receiver operating characteristic curve

auPRC – area under precision recall curve

B[a]P – benzo[a]pyrene

bHLH – basic helix-loop-helix

BPA – bisphenol A

BPS – bisphenol S

ChIP – chromatin immunoprecipitation

CYP – cytochrome P450

DHS – DNase hypersensitive site

DMSO - dimethylsufoxide

DNA – deoxyribonucleic acid

DNase-seq – DNase I hypersensitivity assay followed by sequencing

DRE – dioxin response element

EDC – endocrine disrupting chemical

EGFR – epithelial growth factor receptor

EMSA – electrophoretic mobility shift assay

ERK - extracellular signal-regulated kinases

GD – gestational day

FP – false positive

FPR – false positive rate

GST - glutathione S-transferase

HM – histone modification

IV - intravenous

KLF6 – Krüppel-like factor 6

NF-kB – nuclear factor kB

NTS – nuclear translocation signal

PAH – polycyclic/polyhalogenated aromatic hydrocarbon

PAS – PER-ARNT-SIM

PBTK – physiologically based toxicokinetic modeling

PER - period

PRC – precision recall

QSAR – quantitative structure-activity relationship

ROC – receiver operating characteristic

SIM – single-minded

SP1 – specificity protein 1

SRC - Proto-oncogene tyrosine-protein kinase Src

TAD – transactivation domain

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

TF – transcription factor

TP – true positive

TPR – true positive rate

VC – vehicle control

XRE – xenobiotic response element

XAP2 – hepatitis B virus X-associated protein

CHAPTER 1: INTRODUCTION

Proteins are one of the main building blocks of cells. They are comprised of a chain of smaller units called amino acids that are folded into a functional 3D structure. The precise 3D structure of a protein is crucial to facilitate the performance of its specific cellular functions including enzymatic actions and involvement in providing and maintaining cellular structure (1). Transcription factors (TFs) are proteins that bind to DNA and regulate the transcription of genes, by either promoting or interfering with the recruitment of cellular transcription machinery (2). Some TFs are only activated by a ligand without which they are kept sequestered in the cytosol, and without which they do not bind to DNA nor actively promote transcription (3). Ligand binding determines the activity of such TFs. In that sense, ligands can be both agonist – transforming the TF into an active or DNA binding form; and antagonist – transforming the TF into an inactive form (4).

The aryl hydrocarbon receptor (AhR) is a ligand activated transcription factor belonging to the basic helix-loop-helix (bHLH) PER-ARNT-SIM (PAS) superfamily of TFs which act as sensors of both internal and external cellular environments (5). The existence of the AhR was hypothesized as early as 1976 by Poland et al. (6). Early research on the AhR uncovered its role as a xenobiotic sensor that binds exogenous ligands - xenobiotics in the class of polyhalogenated and polycyclic aromatic hydrocarbons (PAHs), chief among them being 2, 3, 7, 8 tetrachlorodibenzo-*p*-dioxin (TCDD) (7). More recently it has been shown that the AhR can bind endogenous ligands as well. For example, certain tryptophan derivatives – such as kynurenine, tetrapyrroles, and metabolites of arachidonic acid (8–10).

The structure of the AhR protein can be broken down into five main structural domains: the basic domain, two PAS domains, HLH domain and the transactivation domain (TAD). The

basic domain plays a role in DNA binding by the AhR. The two PAS domains - PAS A and PAS B, together with the HLH domain, play a role in dimerization with the canonical AhR dimerization partner – AhR nuclear translocator (ARNT). AhR ligands bind within the PAS B domain. Finally, the C-terminal of AhR contains the transcription activation domain (TAD), which is composed of an acidic, a Q-rich, and P/S/T subdomains. The TAD domain is responsible for the recruitment of co-activators and co-repressors, resulting in the activation or repression of gene expression, respectively (11–13).

Activation of AhR by its exogenous ligands underlies its role as a xenobiotic sensor. The AhR signaling pathway resulting from such activation is referred to as the canonical AhR pathway. Prior to ligand activation, the AhR is localized to the cytosol where it is maintained in its inactive form through binding to its co-chaperone proteins. These proteins include hepatitis B virus X-associated protein (XAP2) – also known as the AhR interacting protein (AIP), a dimer of heat shock protein 90 (HSP90), prostaglandin E synthase 3 (p23), and protein kinase SRC. Upon ligand binding, the AhR is released from its co-chaperones, exposing its nuclear translocation signal (NTS). Subsequently, the AhR translocates to the nucleus where it forms a heterodimer complex with the AhR Nuclear Translocator (ARNT). The AhR-ARNT complex modulates the expression of its target genes by binding to DNA at specific dioxin response elements (DREs), also known as xenobiotic response elements (XREs) and defined by the core DNA sequence 5'-GCGTG-3' (5). This pathway (illustrated in Figure 1) fits within the initially discovered role of AhR as a xenobiotic sensor regulating the adaptive metabolic response. Many target genes in this pathway are xenobiotic metabolizing genes, such as phase I metabolic enzymes, namely cytochrome P450 1A1 (CYP1A1), CYP1A2, CYP1B1, as well as phase II metabolic enzymes, namely glutathione S-transferase (GST), and aldehyde dehydrogenase 3a1 (ALDH3A1) (14, 15). These metabolic

enzymes play a role in detoxifying xenobiotics, but can sometimes produce reactive metabolites, such as in the case of benzo[a]pyrene (B[a]P) (16). On the other hand, the AhR co-chaperone SRC participates in the non-genomic mechanisms of AhR signaling, where its disassociation from the activated AhR can result in the activation of other pathways, namely ERK1/2 and EGFR (17).

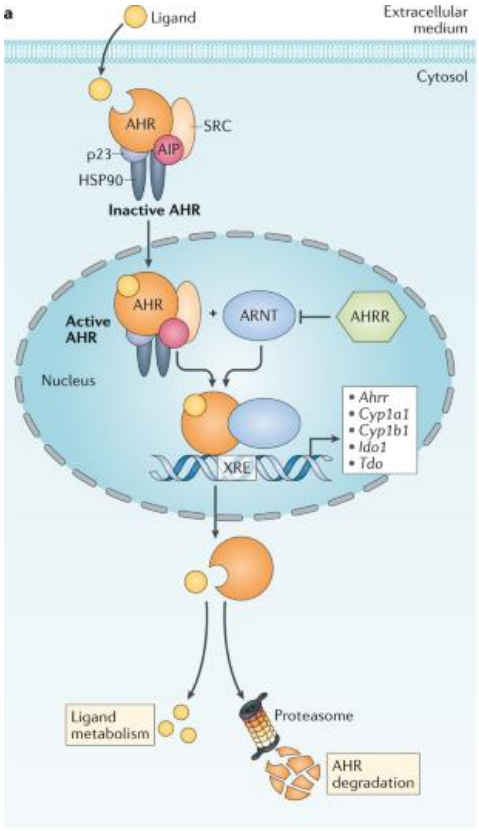

Figure 1 – canonical AhR pathway adapted from (18).

In addition to its role as a xenobiotic sensor, over time many other functions of the AhR have been discovered, such as its roles in differentiation (19), development (20), cancer (21), circadian rhythm (22), cell cycle progression (23), and immunity (24). In general, the AhR can be thought of as an integrator of dietary, metabolic, microbial, and environmental cues that initiates fine-tuned and selective transcriptional programs. These programs can be ligand-specific, cell type-specific and even context-specific (5).

The AhR appears to bind exclusively to DREs in an in vitro setting, i.e., on naked DNA outside of a cell nucleus. When examining the DNA binding of AhR in vivo, by looking at publicly available AhR binding experiments, I observed that most DREs in accessible chromatin are not bound by AhR, and DREs accessible in multiple cell lines or cell types can be bound in some and unbound in others. Nevertheless, the fact that AhR possesses a strong core binding motif 5'-GCGTG-3' – the DRE, facilitates a motif-centered analysis of its binding. To investigate the molecular determinants of AhR binding I developed interpretable machine learning models predicting the binding status of DREs in MCF-7, GM17212, HepG2 cells, and primary human hepatocytes. My results indicate that AhR binding is driven by a complex interplay of cell-agnostic DRE flanking sequence and cell-specific local chromatin context.

Aside from ARNT, AhR has been shown to bind DNA by interacting with other transcription factors, both with and without ARNT. For instance, the AhR-ARNT heterodimer interacts with the specificity protein 1 (SP1) via the AhR-ARNT HLH/PAS domains and SP1 zinc finger domains. AhR-ARNT and SP1 synergistically enhance the transcription of the CYP1A1 gene by binding to their cognate binding motifs in the promoter of CYP1A1. The DRE and the GC-rich binding motif of SP1 in the promoter of CYP1A1 were shown to partially overlap (25). On the other hand, the AhR was shown to interact with several other TFs without ARNT. For example, TCDD activated AhR interacts with the retinoblastoma tumor suppressor protein (pRb) without ARNT to induce G1 cell cycle arrest. The pRb appears to preferentially associate with the ligand-bound form of AhR (26). Additionally, the RelA subunit of nuclear factor-kB (NF-kB) interacts with the AhR to activate the transcription of c-myc and IL-6 genes by binding together in their promoters (27, 28). The transcription of PAI-1, p21cip1, and E-cadherin genes is thought to be driven by AhR interacting with the Krüppel-like factor 6 (KLF6) (29).

However, these ARNT independent AhR-protein interactions have only been confirmed at a limited number of AhR bound loci. It is unknown if these protein interactions with AhR could be occurring genome wide. To address this question, I developed a computational method to investigate AhR-TF interactions at AhR-bound sites. My results confirm ARNT as the main dimerization partner of AhR genome-wide in TCDD-exposed MCF-7 cells. By contrast, in untreated HepG2 cells, KLF6 and RelA but not ARNT were the main dimerization partners of AhR. These findings indicate that the role of AhR is likely ligand-dependent and can potentially be explained through dimerization with different proteins.

Additionally, unrelated to computational modeling of AhR binding and dimerization, I have developed a set of physiologically based toxicokinetic (PBTK) models for bisphenol A and S (BPA and BPS) in pregnant sheep (30). These chemicals are often used in the manufacturing of polycarbonate plastics, epoxy resins, dental sealants, and plastic and paper consumer products (31, 32). They are also known endocrine disruptors and can be found pervasively in the environment (32, 33). The PBTK models I developed and calibrated against available toxicokinetic data (34–36), demonstrated that BPS exhibited a higher potential for accumulation in the fetus with repeated daily maternal exposure.

CHAPTER 2: AHR BINDING PREDICTION

INTRODUCTION

The expression of genes is governed principally through the process of transcriptional regulation. This process represents the main mechanism by which crucial cellular processes such as differentiation, development, and response to exogenous stimuli are coordinated (37). Transcriptional regulation occurs in large part through direct or indirect binding of transcription factors (TFs) to DNA, and through the interactions of these TFs with transcriptional machinery (38). Altering the expression, function, or the DNA-binding ability of even a single TF can result in changes in expression of hundreds to thousands of genes (39, 40). Further, the removal of a single TF binding site, e.g., through an experimental procedure such as promoter bashing or targeted mutagenesis, often results in altered gene expression (41, 42). The problem of experimentally identifying TF binding sites across the genome is further complicated by the fact that TF binding is often highly tissue- and cell type- specific (37).

The problem of computationally predicting DNA binding sites genome-wide for a particular TF, in a cell type- or tissue-specific manner, can be likened to finding the proverbial needle in a haystack. This problem is particularly difficult for TFs with a short core DNA binding motif such as the aryl hydrocarbon receptor (AhR). The AhR binds to a core 5-base pair (bp) sequence, 5'-GCGTG-3', referred to as the dioxin response element (DRE) (43, 44). Such short binding motifs occur millions of times in the human genome. However, a typical chromatin immunoprecipitation followed by sequencing (ChIP-seq) binding assay for AhR produces a list of AhR bound regions on the order of a few hundred to a few thousand regions. This discrepancy could be partially explained by the fact that most genomic DREs lie in inaccessible regions of the genome in a particular cell line or type. Additionally, the nucleotides flanking the core motif on

both 5' and 3' ends are suspected to form an extended active AhR binding site (45). However, the manner in which the exact identity of these nucleotides affects AhR binding is currently unknown. Further, these AhR bound regions have anywhere between zero and 29 DREs (46, 47). Therefore, the occurrence of a DRE in the genome is neither sufficient nor necessary to induce AhR binding. The problem of predicting AhR-DNA binding is additionally complicated by the fact that it can be hard to distinguish direct DNA binding from indirect binding through tethering with other TFs or through 3D looping of chromatin (48). Considered together, these findings indicate that both genomic and epigenomic characteristics likely play a role in determining AhR binding in vivo.

The AhR is a ligand-activated transcription factor (TF) in the basic-helix-loop-helix (bHLH) PER-ARNT-SIM (PAS) family of TFs (49, 50). The AhR can be activated by both endogenous and exogenous ligands (5, 51). In the later class of ligands, the environmental pollutant 2, 3, 7, 8 tetrachlorodibenzo-ρ-dioxin (TCDD) is the prototypical AhR ligand (7, 51). Exposure to TCDD activates the xenobiotic response pathway of AhR. Initially, the AhR is constrained in the cytosol of the cell through binding with its co-chaperone proteins. These proteins include a dimer of the 90-kDa heat shock protein (HSP90) (52, 53), the AhR-interacting protein (AIP) (54, 55), the cochaperone protein p23 (56), and SRC (17). When bound by its ligand, AhR releases from its co-chaperones and translocates to the nucleus where it forms a heterodimer with the AhR Nuclear Translocator (ARNT) (57, 58). The AhR-ARNT heterodimer binds to DNA sequences containing the consensus 5'-GCGTG-3' core binding motif (59, 60). This binding motif has been named variously the xenobiotic response element (XRE), aryl hydrocarbon response element (AhRE), or dioxin response element (DRE) (61). In this thesis I will use the name DRE when to referring to AhR binding motif within potential AhR DNA binding sites.

The best understood function of the AhR is the direct regulation of its target genes, chief among them the cytochrome p450 1A1 (CYP1A1), 1A2 (CYP1A2) and 1B1 (CYP1B1). The AhR regulates these genes by binding to DREs in their proximal promoters or, potentially, distal enhancers (24, 62, 63). The first step towards reconstructing the AhR-mediated gene regulatory network is the accurate, cell type-specific identification of AhR binding sites. The construction of these gene regulatory networks is crucial for improving our understanding of the role of the AhR in xenobiotic-induced toxicity and disease, as well as in crucial physiological functions. These include the immune response (24), circadian rhythm (22), cell cycle progression (23), and embryonic development (20). Significant progress has been made with the development of high throughput molecular techniques for identification of TF-bound DNA fragments. These techniques often use a method for the enrichment of TF bound DNA complexes followed by sequencing of the enriched DNA fragments. Techniques such as ChIP-seq (64), ChIP-exo (65) and ChIP-nexus (66) have enabled a genome-wide view of TF binding. Over time, the binding of hundreds of TFs in multiple cell lines, primary cells, and whole tissues has been investigated genome-wide and the results of these experiments have been made publicly available. Likewise, DNA binding of the AhR has been probed in several human cell lines and primary cells. Nonetheless, the determinants of cell-specificity of AhR binding remain poorly understood.

Recent years have seen the development of many computational approaches for genome-wide prediction of TF binding. The most widely used methods leverage the position weight matrix (PWM) corresponding to the TF of interest. A PWM is a statistical and quantitative representation of known and experimentally confirmed DNA binding sites for a TF of interest. The PWM effectively makes up the binding motif of the TF. PWMs are available in online databases such as TRANSFAC and JASPAR. These PWMs have been derived from experimental data and can also

be estimated de novo if binding data is available (67, 68). A PWM is used to calculate a score for each potential binding site as a sum of individual scores of each nucleotide making up the PWM and overlapping the potential binding site. PWMs are then used to scan the genome for TF binding sites, using a previously derived optimal threshold score as the cutoff to predict TF-bound sites (69, 70). PWMs are commonly derived from in vitro experiments. The most often used in vitro experiment is the high throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) (71). Occasionally, PWMs are also derived from in vivo experiments such as ChIP-seq. However, when examining the binding of TFs in vivo, it is often noted that many TFs are bound to DNA sequences that do not possess the in vitro or even the in vivo derived binding motif (72).

Eukaryotic TFs generally do not bind DNA in isolation but rather in dense, often tissue-specific, TF clusters. These clusters are characterized by the co-location of the TF binding sites for multiple different types of TFs in relatively short genomic regions (73, 74). Consequently, it is reasonable to assume that PWMs of co-bound TFs could be used to predict the binding of a TF of interest. Still, models that use PWMs of co-binding TFs have shown limited utility in improving model performance, with models of certain TFs seeing little to no improvement (72). Even so, given that TFs bind in dense clusters and that PWMs are not always representative of actual TF binding, I hypothesized that ChIP-seq signals of co-bound TFs, as a measure of their actual binding, could provide the information that PWMs could not. Further, I propose that interpretable machine learning combined with the measures of co-bound TFs would provide mechanistic insights into the molecular mechanisms underlying the cell specificity of AhR binding.

TF binding prediction models based on PWMs have been extended over time to include other biological features demonstrated to be associated with TF binding, such as chromatin accessibility, histone modifications, evolutionary sequence conservation, PWMs of co-bound TFs,

and gene expression (72, 75). Similarly, a broad range of statistical and machine learning models ranging from unsupervised Bayesian mixture models (75) to deep learning (76–78) have been used to address the problem of tissue-specific TF binding prediction. Despite some of these models achieving high cross-tissue performance for select few TFs, most of them lack interpretability and do not translate into mechanistic insights.

Most computational models predicting TF binding have been applied to constitutively active TFs. The binding of inducible TFs, on the other hand, remains largely computationally unexplored. In this chapter of the thesis, I applied a supervised machine learning algorithm, XGBoost (79), and developed machine learning models predicting the AhR binding status of DREs in open chromatin of a particular cell line or type. These models were trained to predict DREs in open chromatin, as either bound or unbound, and were applied to four cell lines and one primary cell type: two human breast cancer cell lines (MCF-7 and T-47D) (46, 47), primary human hepatocytes (80), human hepatocellular carcinoma cell line (HepG2) – data obtained from the ENCODE project (81, 82), and lymphoblastoid cell line (GM17212) (83). The cells in these experiments were treated with either TCDD, Methylcholanthrene (3-MC; an AhR ligand) or Dimethyl sulfoxide (DMSO; vehicle control) for a duration of either 45 minutes, 1 hour or 24 hours. By using these datasets and chromatin accessibility experiments corresponding to the cell line or type used, I first identified cell line and type-specific AhR- bound and unbound DREs in open chromatin. Then, I developed machine learning models that predict the binding status of DREs in open chromatin for each cell line or type individually. My results demonstrate highly accurate and robust models of within-cell line or cell type binding. I identified several TFs as predictive of AhR binding in individual cell lines or types, such as GATA3 in MCF-7 cells, MXI1 in HepG2 cells, and SP1 in primary human hepatocytes and GM17212 cells; as well as histone

modifications (HMs) – H3K4me1 and H3K4me3 in MCF-7 cells, H3K4me3 and H3K27ac in primary hepatocytes, and H3K27ac in GM17212 cells. My cell-specific models generalize well to the prediction of AhR binding sites without DREs, demonstrating the robustness of the models. In conclusion, I demonstrated that the patterns of TFs and HMs most predictive of AhR binding are consistent within cell lines or types but highly variable across them, which is suggestive of potentially different underlying cell-specific mechanisms of AhR binding. Additionally, I show that AhR binding is driven by a complex interplay of cell-agnostic DNA sequence flanking the DRE and cell-specific local chromatin context. The approach used here can be adapted to other inducible TFs, such as steroid hormone and nuclear receptors.

MATERIALS AND METHODS

Reference genome

Unless otherwise specified the reference, genome used for sequence alignment in this part of the thesis was the human genome assembly version hg19. I opted for hg19 due to availability of data on ChIP-seq and DNase-seq data repositories such as GEO Datasets (84) ChIP-Atlas (85) and ENCODE (81, 82). Likewise, most other transcription factor binding prediction tools available at this time were trained on hg19.

Visualization of ChIP-seq signal

Bigwig files were used as inputs to deepTools version 3.5.1 (86) for visualization. DeepTools plotHeatmap function was used to create visualizations of ChIP-seq signal fold enrichment within a −1.5 to +1.5 kb region around the bound and unbound dioxin response elements (DREs), as well as to generate average profiles for ChIP-seq enrichment in the same region.

DREs in open chromatin

I obtained DNase-seq data for all relevant cell lines (MCF-7, T-47D, primary hepatocytes, HepG2, GM12878) from ENCODE - https://encodeproject.org/. I downloaded the broadPeak DNase-seq files for the hg19 genome assembly, and if there were multiple replicates, I found the intersection of all replicates. Any DRE found under the peaks of DNase-seq intersection was considered to be in the open chromatin of the corresponding cell line and was used in the determination of bound and unbound DREs for the purposes of model training. DREs occurring in ENCODE blacklisted regions, namely the merged consensus blacklist (wgEncodeHg19ConsensusSignalArtifactRegions.bed.gz) and exclusion list regions (ENCODE accession ENCFF001TDO) were ignored in downstream analyses.

AhR-bound and unbound DREs

Firstly, I assembled a list of all DREs in the human genome by searching the hg19 human reference genome sequence for the occurrences of the core DRE sequence 5'-GCGTG-3' on either strand of the DNA. Only DREs in open chromatin, i.e., DREs overlapping DNase-seq broadPeaks from an ENCODE experiment for a given cell line, were considered for training. Additionally, bound DREs in closed chromatin were also considered for testing purposes. Secondly, I obtained the AhR ChIP-seq bed and bigwig files either from Gene Expression Omnibus (GEO) Datasets or from ChIP-Atlas (85) where the original sequencing files have been processed uniformly following a standard processing pipeline. Originally, AhR ChIP-seq data was generated in the following independent experiments, 1) AhR and ARNT ChIP-seq of MCF-7 cells treated with 10 nM TCDD for 45 minutes (47) – the binding data was obtained from GEO Datasets – accession GSE41820, 2) AhR and AhRR ChIP-seq of MCF-7 cells treated with 10 nM TCDD for 45 minutes and 24 hours (46) - the binding data was obtained from GEO Datasets – accession GSE90550, 3) AhR

ChIP-chip of T-47D cells treated with 1 μM 3-MC or 10 nM TCDD for 1 hour (87, 88) - the binding data was obtained from their respective publications and converted from hg18 to hg19 using the liftOver tool (89) 4) AhR ChIP-seq of primary hepatocytes treated with 1nM of TCDD for 24 hours – the binding data was obtained from GEO Datasets – accession GSE205502; 5) AhR (3xFLAG tagged AhR) ChIP-seq of untreated HepG2 cells from ENCODE – accession ENCSR412ZDC (81, 82); 6) AhR ChIP-seq of GM17212 cells treated with 1 μM 3-MC for 24 hours – accessible through GEO Datasets – accession GSE116632; however, the binding data was obtained from ChIP-Atlas – accessions SRX4342282, SRX4342283, SRX4342285, and SRX4342286. Details of these experiments are summarized in Table 1. Bound DREs for the purposes of model training were determined as DREs found in open chromatin and under AhR peaks where only one DRE was present under the AhR peak (referred to as singleton DREs). Isolated unbound DREs are DREs in open chromatin found at least 500 bps away from the boundary of any AhR peak, as well as 100 bps away from any other DRE. These DREs were selected as unbound for model training in order to minimize confounding of DRE contribution to binding. All other DREs in open chromatin were considered ambiguous and were not used in model training.

Promoters and enhancers

I obtained all annotated transcription start sites from Ensembl 105 BioMart (human genes; GRCh38.p13) (90) and considered regions ±200 and ±1500 bp around the TSS as stringent and relaxed promoters, respectively. I obtained all computationally predicted enhancers from ChromHMM (91) for samples that had ChromHMM data available – HepG2 and GM12878 (ENCODE) and MCF-7 (GEO Datasets – accession GSE57498). Both weak and strong enhancers (ChromHMM states 4 through 7) were considered as valid enhancers.

Sequence and genomic signal features

For each DRE in the human genome, I obtained the genomic sequence of seven nucleotides 5' upstream and 3' downstream from the DRE (5'-GCGTG-3') from the hg19 human reference genome. These nucleotides were one-hot encoded and used as features in my machine learning models. In total there were around 1.6 million DREs spread across the human genome. However, only a small fraction of them fulfilled the criteria for bound and unbound DREs used in training and testing. DNase-seq, as well as all available histone mark and transcription factor ChIP-seq genomic signal (bigwig) files were downloaded for MCF-7, T-47D, primary hepatocytes, HepG2, GM12878 (as the closest match to GM17212 where AhR was ChIP-ed) from the ENCODE consortium. For each bound and unbound DRE and each genomic signal (bigwig) file, I extracted the value of the genomic signal 740 bps up- and 740 bps down- stream from the DRE, for a total of 1485 bps of signal (DRE width is 5 base pairs). The extracted signal was split into 15 bins of equal 99-bp size and the signal within each bin was averaged to produce 15 features corresponding to the particular DRE-genomic signal combination. During averaging, any areas of missing signal were replaced with zeros.

Model architecture and training

For each cell line and all the bound and unbound DREs appearing in open chromatin of that particular cell line, I created sequence features, as well as genomic signal features for all available DNase-seq, histone mark and transcription factor (TF) ChIP-seq experiments. I then performed hyperparameter tuning of an XGBoost model through a grid search of the hyperparameter space with the following values - max_depth = {3, 4, 5, 6, 7}, min_child_weight = {3, 4, 5, 6, 7}, subsample = {1.0, 0.9, 0.8, 0.7}, colsample_by_tree = {1.0, 0.9, 0.8, 0.7} and eta

= {0.05, 0.075, 0.1, 0.125, 0.15, 0.2, 0.3}. I reported the average performances over all five folds for the best performing models in terms of hyperparameter selection.

Model evaluation

In addition to evaluating the models through 5-fold cross validation, I also evaluated model performance on predicting the binding status of DREs that occurred under multi-DRE AhR peaks both in open and closed chromatin. For each such peak and each DRE under the peak I used the AhR binding prediction model to make a prediction regarding whether the DRE is bound or not. If at least one DRE under the peak was predicted as bound, the peak was considered recovered, and the total fraction of recovered peaks was reported. Similarly, I evaluated the model performance on predicting the binding status of AhR peaks without DREs. Briefly, for each 0-DRE peak I simulated five dummy DREs. These dummy DREs are not actually present in the genomic sequence and only represent the genomic location that was used as reference for the calculation of all non-sequence model input features. The center of the first dummy DRE is aligned to the center of the AhR peak and the other four dummy DREs are positioned -100, -50, +50, +100 bps relative to the center point of the first dummy DRE. A zero-DRE peak is considered generally recovered if at least one of the five dummy DREs is predicted as bound. The peak is considered centrally recovered if the central DRE is predicted as bound.

Model performance metrics

To calculate the area under Receiver Operating Characteristic (auROC) and area under Precision Recall (auPRC) curves, I used the output of the XGBoost algorithm in the form of probabilities of each particular observation (DRE) belonging to a particular output class (bound or unbound). By using different thresholds for these probabilities above which the model predicts a DRE as bound, I obtained the numbers of true and false positives for each threshold, as well as

true and false negatives relative to the ground truth of DRE binding obtained from the corresponding AhR ChIP-seq experiment. Each threshold produced a point on the ROC and PRC curves; the area under the curve was calculated using a line interpolated through all the points.

Statistical analysis

Statistical analysis was carried out in Python 3, using the scipy 1.8.0 package (92). ChIP-seq signals were analyzed using the Kruskal-Wallis test and post-hoc analysis performed with the Wilcoxon test for each pair. Results were considered significant if P-value was < 0.01.

RESULTS

AhR binding is cell line and cell type-specific

The first part of my study was focused on improving the understanding of the molecular determinants underlying the binding of the human aryl hydrocarbon receptor (AhR) to DNA. To achieve this goal, I investigated the role of the core 5'-GCGTG-3' AhR binding motif in determining the cell-specificity of AhR binding. This core AhR binding motif is known as the dioxin response element – DRE. I compared AhR binding in human cells across previously published and publicly available AhR binding data in the form of chromatin immunoprecipitation (ChIP) experiments. These experiments were either followed by sequencing (ChIP-seq) or a microarray (ChIP-chip). The ChIP-seq experiments provided a genome-wide view of AhR binding, while the ChIP-chip experiments were focused only on determining AhR binding in gene promoters. Each experiment selected for the analysis of AhR binding was performed on a specific cell line or on primary cells. Experiments on the following cells were included in further analyses – 1) two epithelial breast cancer cell lines - MCF-7 and T-47D, 2) a hepatocellular carcinoma cell line - HepG2, 3) a lymphoblastoid cell line – GM17212, and 4) primary human hepatocytes. The cells in these experiments were treated with an AhR agonist (TCDD or 3-MC) or vehicle control

(VC) – dimethyl sulfoxide (DMSO) for a duration of 45 minutes, 1 hour or 24 hours, or were not treated with either – HepG2 cells. For the full list of experiments, including the total number of AhR peaks, as well as type, concentration, and duration of treatments, see Table 1. The breast cancer cell lines had data available from more than one experiment.

	Treatment						
Cells	Chemical	Duration	Concentration	Control	Genome wide?	Number of peaks	GEO or ENCODE accession
MCF-7	TCDD	45 minutes	10 nM	TCDD+IgG	Yes	2594	GSE41820
MCF-7	TCDD	24 hours	10 nM	VC	Yes	3494	GSE90550
primary hepatocytes	TCDD	24 hours	1 nM	VC	Yes	3145	GSE205502
HepG2	none	0 hours	none	IgG	Yes	12164	ENCSR412ZDC
GM17212	3-MC	24 hours	1 µM	VC	Yes	17535	GSE116632
T-47D	3-MC	1 hour	1 µM	VC	No	241	None
T-47D	TCDD	1 hour	10 nM	VC	No	411	None

Table 1 – AhR ChIP-seq and ChIP-chip experiments.

To develop and train my machine learning models I used data from four AhR ChIP-seq binding experiments - 1) MCF-7 cells treated with 10 nM TCDD for 24 hours (referred to as MCF-7 or MCF-7 24h), 2) primary hepatocytes treated with 1 nM TCDD for 24 hours (referred to as primary hepatocytes), 3) HepG2 cells without AhR agonist treatment (referred to as HepG2), 4) GM17212 cells treated with 1 µM 3-MC for 24 hours (referred to as GM17212). In addition, some analyses were performed on the remaining three AhR ChIP-seq and ChIP-chip binding experiments – 1) MCF-7 cells treated with 10 nM TCDD for 45 minutes (referred to as MCF-7 45m), 2) T-47D cells treated with 10 nM TCDD for 1 hour (referred to as T-47D TCDD), 3) T-47D cells treated with 1 µM 3-MC for 1 hour (referred to as T-47D 3-MC). All AhR binding

experiments performed on T-47D cells were ChIP-chip and only reported the binding of AhR in gene promoters.

First, I searched for the occurrences of DREs within the hg19 human reference genome and found approximately 1.6 million DREs in the human genome. Upon intersecting these DREs with the genomic locations of AhR peaks I identified the existence of AhR peaks with either (i) none (0-DRE peaks), (ii) exactly one (singleton peaks), or (iii) more than one (multi-DRE peaks) DREs (Figure 2).

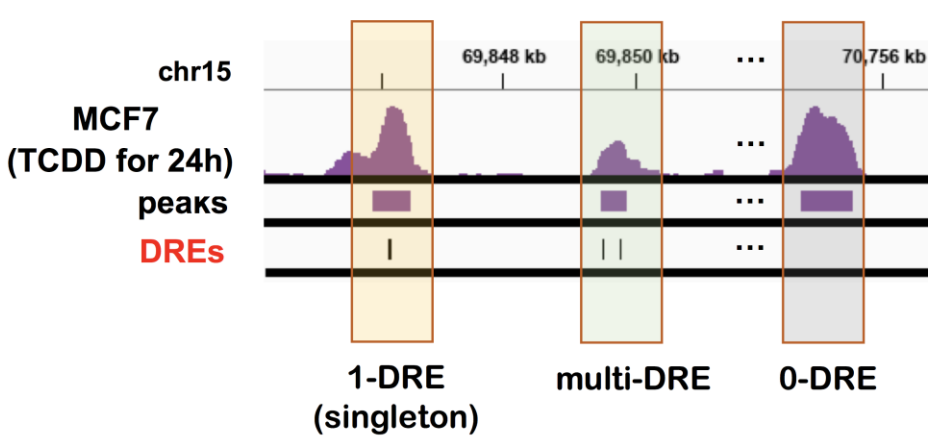

Figure 2 – AhR peaks with 0-, 1-, and multi-DREs under the peak.

Next, I calculated the percentages of each type of peak – 0-DRE, singleton, and multi-DRE, across all AhR peaks within each AhR binding data set listed in Table 1. The percentage of singleton AhR peaks ranged between 22.3% and 33% and was similar across all data sets. The percentage of multi-DRE peaks, however, was markedly larger for the two types of liver cells – HepG2 cells and primary hepatocytes (Figure 3). Even though HepG2 cells were not treated with an AhR agonist, the HepG2 experiment resulted in 12,164 AhR peaks. These results suggest the possibility of basal induction of AhR in cells exposed to typical cell culture conditions.

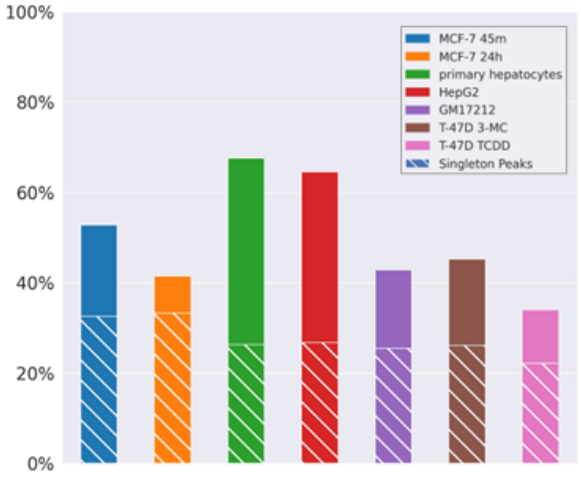

Figure 3 – percentage of AhR peaks with at least one DRE.

Then, I investigated the locations of singleton DREs relative to the mid-point of their corresponding AhR peaks. If these DREs are indeed functional and not only showing up at random under the AhR peaks I would expect them to be enriched at one central or two centrally symmetrical points. This is because the AhR-ARNT dimer can bind to either strand of the DNA – depending on where the DRE is located, and because the DNA binding domain of the dimer does not lie exactly in the middle of the protein complex. My results show that the majority of singleton DREs are located near the mid-point of their corresponding AhR peak (Figure 4). For example, in MCF-7 cells, approximately 50% and 80% of singleton peak DREs were found within 100 and 200 base pairs up-/down- stream from the midpoint of the peak, respectively. However, I observed that in HepG2 cells, even though singleton DREs appeared somewhat centrally enriched, it was not to the same degree as singleton DREs in MCF-7 cells, primary hepatocytes and GM17212 cells (Figure 4).

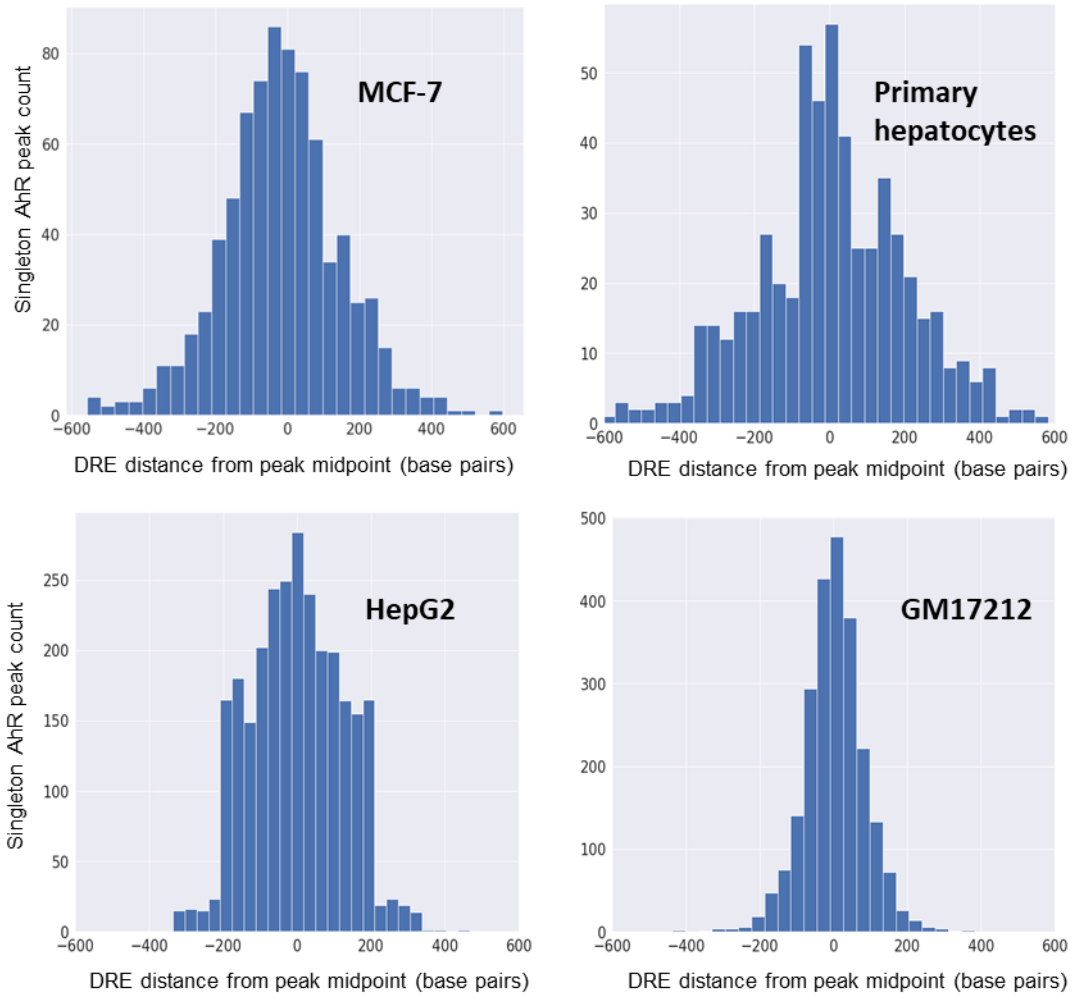

Figure 4 – histogram of DRE position relative to the mid-poing of singleton AhR peaks.

Most TF binding peaks occur in areas of open chromatin, except for some pioneering factors that bind in areas of closed chromatin, and potentially recruit chromatin remodeling factors. Therefore, the TF binding peaks of most TFs overlap with DNase hypersensitive sites (DHSs) or similar experimentally verified regions of open chromatin (82). To examine the contribution of chromatin accessibility in determining cell specificity of AhR binding, I examined DNase-seq broadPeak files from the ENCODE database. Only DNase-seq experiments most closely corresponding to the cell line or primary cell type used in the AhR binding experiment were used (81, 82). I identified all DREs and AhR peaks appearing in open chromatin for each cell line or type and determined that the majority of AhR peaks can be found in open chromatin (Table 2).

Cells	Number of DREs in open chromatin		Percentage of AhR peaks in open chromatin
	Singleton bound	Isolated unbound	
MCF-7	869	27 486	92.9%
primary hepatocytes	764	39 750	69.0%
HepG2	3 857	21 228	83.6%
GM17212	2 557	3 833	85.6%

Table 2 – number of singleton bound and isolated unbound DREs; percentage of AhR peaks in open chromatin.

Generally, DNase-seq or other types of experiments probing chromatin accessibility available in databases such as ENCODE are done on cells under normal cell culture conditions. Consequently, none of the available DNase-seq experiments corresponding to the cell lines or types used in this thesis were treated with AhR agonists. Therefore, the DNase-seq and AhR binding experiments are matched in ligand treatment only in the HepG2 cells, since these cells were not explicitly exposed to an AhR ligand. However, even though I use DNase-seq experiments corresponding to a non-treated cellular state to determine bound and unbound DREs in a treated cellular state, I observed that the majority of AhR peaks do lie in open chromatin - as it is before treatment (Table 2) – between 83.6% and 92.9%. These findings are consistent with our ATAC-seq data in mouse primary hepatocytes treated with TCDD for 6 hours (unpublished). These results indicate that AhR activation does not result in extensive chromatin remodeling, therefore the use of the DNase-seq prior to treatment is justified. The only exception I found was the human primary hepatocyte AhR ChIP-seq experiment where only about 69% of AhR peaks lie in initially open chromatin. Unfortunately, none of the related human liver or hepatocyte DNase-seq experiments available on ENCODE could be closely matched to the primary hepatocytes used for the AhR

ChIP-seq experiment. Hence, primary hepatocytes were excluded from many of the subsequent analyses.

Further analysis focused on contrasting bound and unbound DREs. I observed that the proportions and exact identities of unbound and bound DREs found in open chromatin appear to be highly cell line or type specific. Out of nearly 8,000 DREs found in open chromatin across each of the four relevant cell lines or types, about half are bound in at least one, while only 14 DREs are bound in all four (Figure 5).

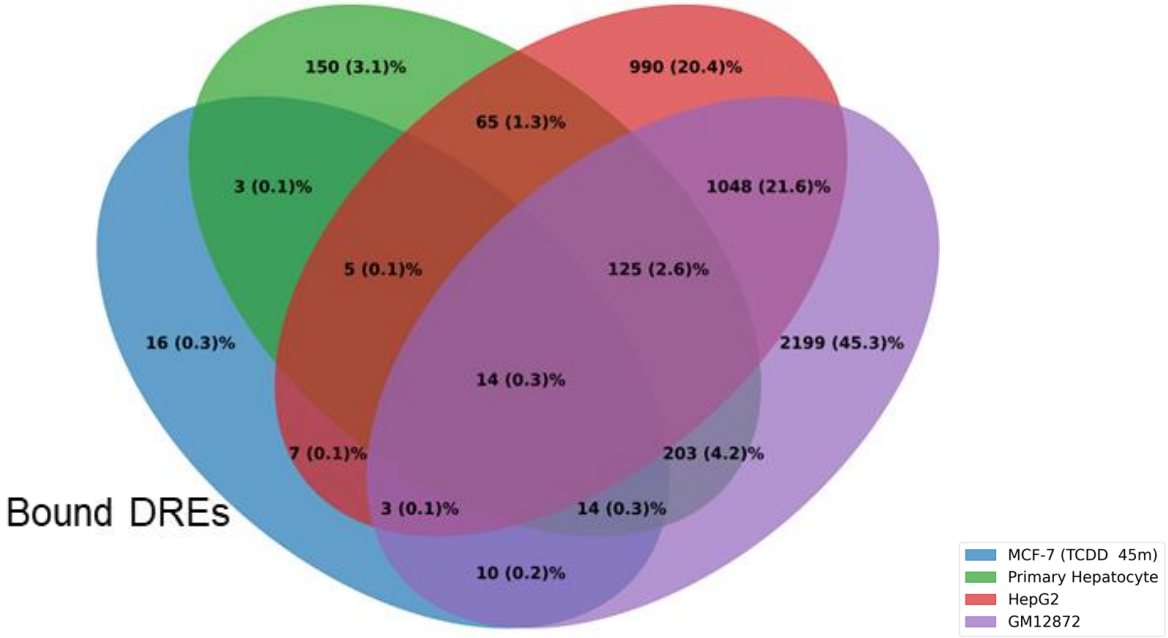

Figure 5 – Venn diagram of bound DREs in accessible chromatin of all four cell lines or types.

In contrast, about half of these pervasively accessible DREs are unbound in all four cell lines or types (Figure 6). These results suggest that if a DRE is found in open chromatin of all four cell lines or types that DRE is much more likely to be unbound in all four than it is to be bound. Conversely, if such a DRE is found to be at all bound than it is most likely bound in only one or at most cell lines or types, since only 3.1% of bound DREs in Figure 5 are bound in 3 or all 4 cell lines or types.

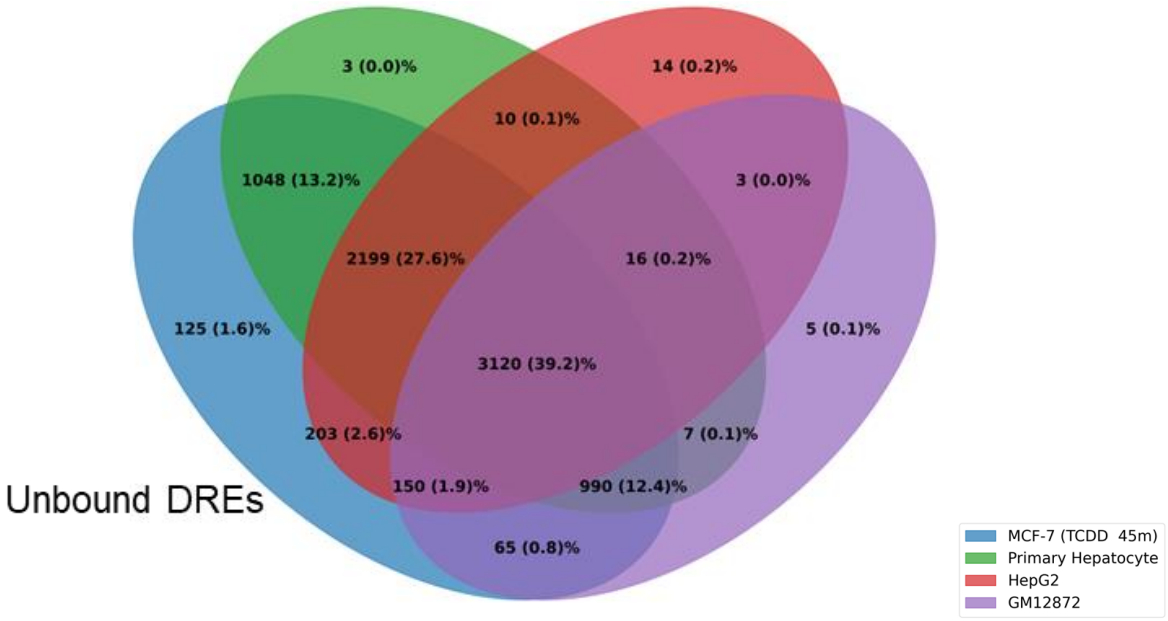

Figure 6 – Venn diagram of unbound DREs in accessible chromatin of all four cell lines or types.

The two breast cancer cell lines, MCF-7, and T-47D, have publicly available AhR binding experiments with similar treatment conditions, i.e., 45 minutes of 10 nM TCDD and 1 hour of 10 nM TCDD treatment, respectively. However, even between these two breast cancer cell lines most accessible DREs appear bound in only one of the two cell lines according to their respective AhR peak lists (Figure 7). A more detailed look at a heatmap of MCF-7 AhR binding signal strength surrounding DREs bound only in T-47D shows that as many as three quarters of these DREs also possess subthreshold peaks in MCF-7 cells (Figure 7C). Ultimately, a DRE that lies within open chromatin of two different cell lines or types is somewhat likely to be bound in one and unbound in the other. These results jointly suggest the existence of AhR binding determinants beyond DNA sequence and the accessibility of chromatin.

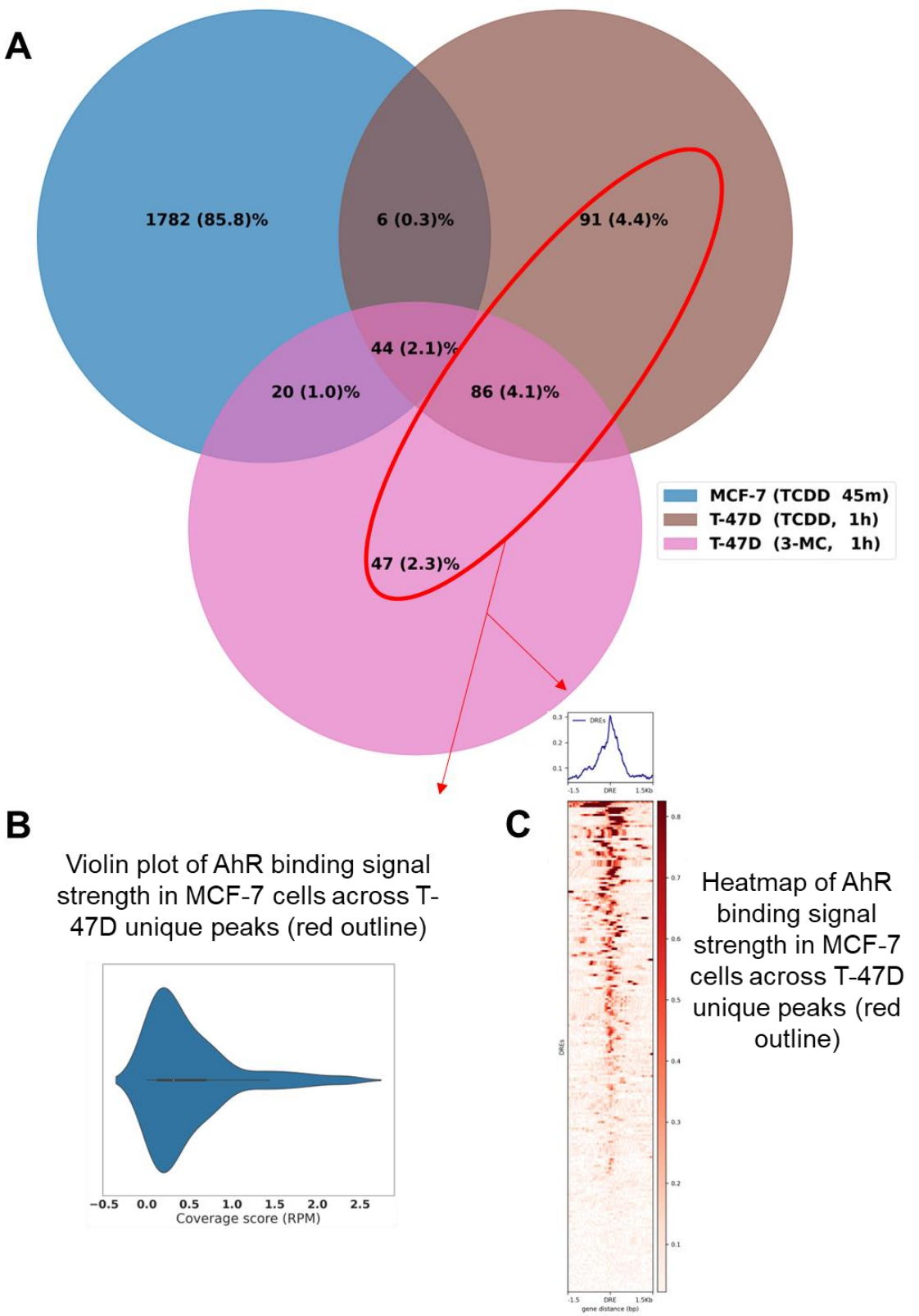

Figure 7 – Venn diagram of AhR peaks in MCF-7 and T-47D cells; MCF-7 AhR signal across T-47D only peaks.

When examining AhR peaks shared between two cell lines or types, I observed that among AhR peaks found at the same genomic location in two binding experiments there was a higher percentage of peaks with DREs, when compared to AhR peaks that were unique to a single binding experiment (Figure 8). Conversely, this means that AhR peaks with DREs are more likely to appear in more than one cell line or cell type than 0-DREs peaks.

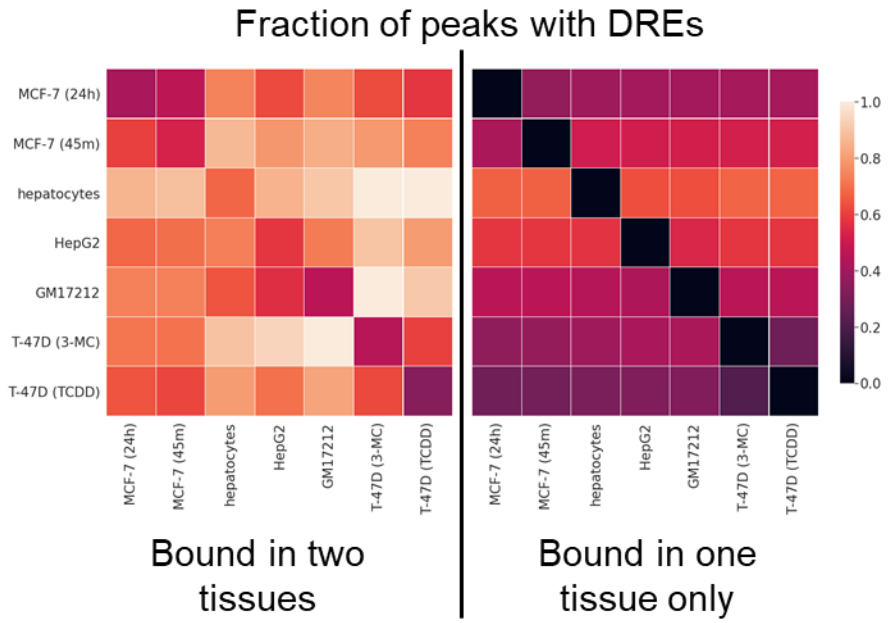

Figure 8 – fraction of AhR peaks with DREs.

Breaking down AhR peaks by the number of DREs under the peak, into 0-DRE, singleton, and multi-DRE peaks and then quantifying the AhR ChIP-seq signal across each group of peaks revealed significant differences between average signal strength between groups. Namely, the more DREs an AhR peak had, the higher the average ChIP-seq signal was under the peak (Figure 9). These results jointly suggest that DREs are likely participating in determination of AhR binding and that a DRE-centric approach to the investigation of cell-specificity of AhR binding could reveal important determinants and potential mechanisms driving AhR binding to DNA.

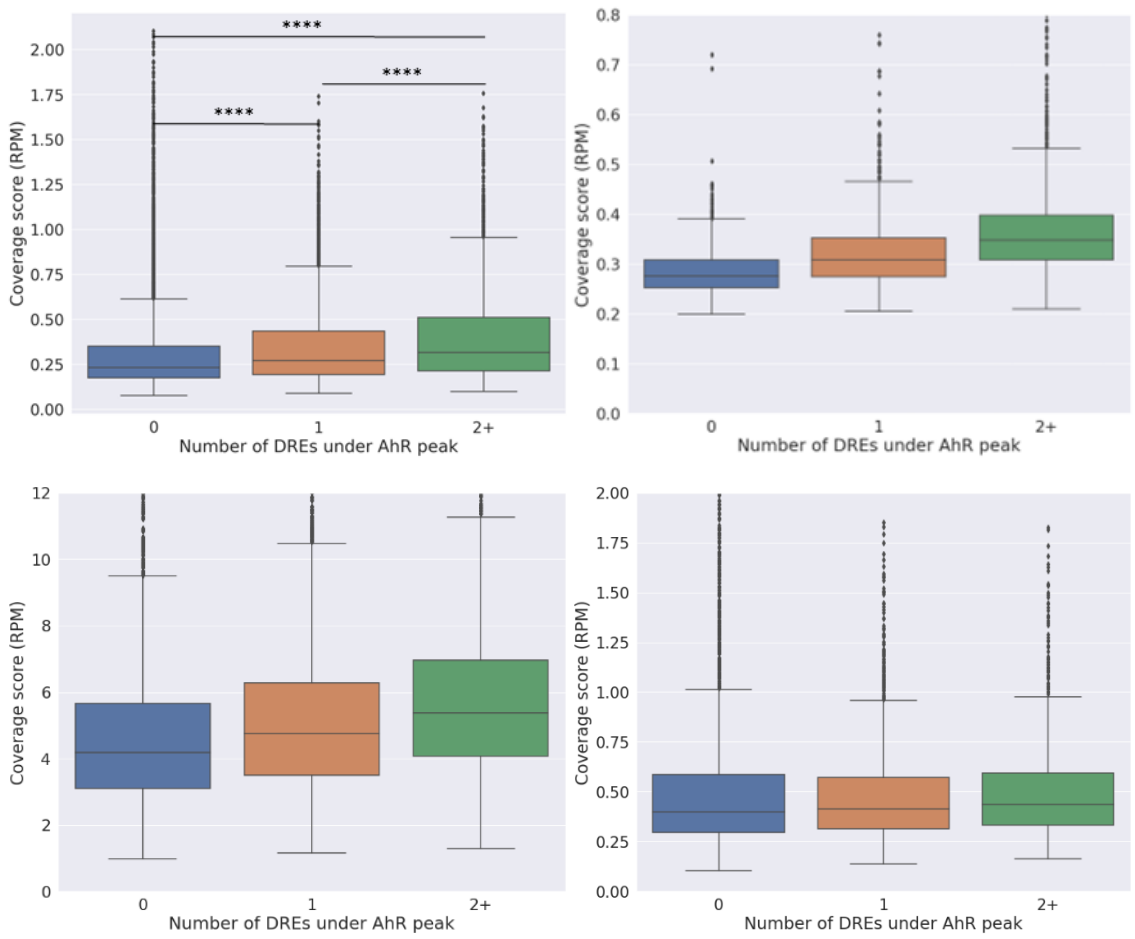

Figure 9 – average AhR signal across peaks with 0-, 1-, and multi-DREs.

Machine learning models accurately predict AhR binding

Next, I sought to improve our understanding of the likely molecular determinants of cell-specific AhR binding, beyond chromatin accessibility and the core DRE motif. To achieve this goal, I developed a set of interpretable machine learning models trained to predict the binary binding status of DREs in open chromatin, i.e., bound or unbound. The models were trained with increasingly complex combinations of input features for each cell line or type. The models were trained on the singleton bound DREs occurring under AhR peaks, as my bound (positive) training examples and isolated unbound DREs (see Methods) occurring in open chromatin but not under AhR peaks, as my unbound (negative) training examples. The DREs under multi-DRE peaks were considered ambiguous, as it was not possible to computationally determine which specific DRE(s)

among the cluster of DREs were responsible for AhR binding. However, I have used these DREs in model evaluation. All machine learning models presented in this thesis were developed using the gradient boosted tree algorithm of the XGBoost family of algorithms, which has been shown to handle non-linear data well (93). In addition, these algorithms also provide metrics of feature importance. Therefore, it is possible to evaluate the contribution of individual input features to improving the model performance (94).

The models I developed use features centered on the DRE which were based on the local chromatin context. These models are trained on singleton bound and isolated unbound DREs found in open chromatin for the cell line or type of interest. Models were validated using the 5-fold cross validation procedure. Due to a limited number of bound singleton DREs, I have not created a dedicated test set to evaluate the models. Instead, as model hyperparameters are tuned with 5-fold cross validation (see Methods), the average performance of the models across the five folds is reported (Figure 10). This choice is further justified by the purpose of this thesis which was to create interpretable machine learning models and derive from them mechanistic insights regarding cell-specific binding of AhR.

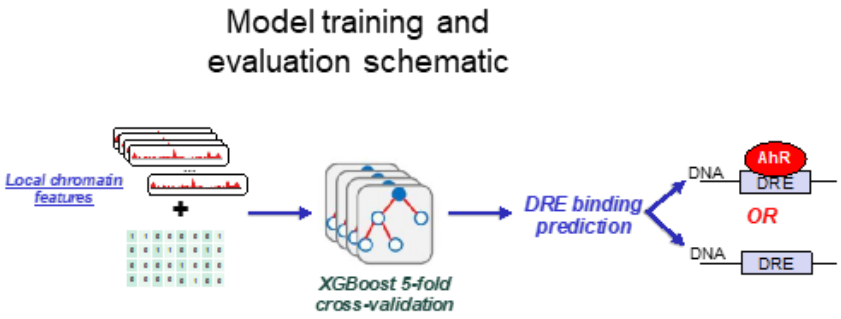

Figure 10 – schematic representation of model training.

I included the following local chromatin input features in the trained models - 1) DNA sequence immediately flanking the DRE. The contribution of flanking sequence of up to 7 nucleotides directly up- and down- stream from the DRE was investigated. These nucleotides have

been previously proposed to be involved in AhR binding through an analysis of 13 bona fide AhR binding sites (45). The flanking sequences were one-hot encoded and used as model inputs; 2) Binned average values of bigWig signals of experiments performed on the cell line or type most closely corresponding to the one used in the AhR binding experiment. Namely, for the primary hepatocyte model I used bigWig signals from experiments done in hepatocytes originated from H9 cells, and for GM17212 I used bigWig signals from experiments done in GM12878. All other cell lines were matched exactly, e.g., MCF-7, and HepG2. To create model input features from these bigWig files I used the following publicly available sequencing experiments i) DNase-seq (as representative of chromatin accessibility), ii) histone modification, and iii) transcription factor ChIP-seq experiments from ENCODE – see methods for details (81, 82). I created 15 bins of width 99 base pairs for each bigWig signal and each DRE. Each bin was assigned a value that was the average bigWig signal across the width of that bin. The mid-point of the central bin was positioned at the middle nucleotide of the 5-bp DRE; 3) Indicator variables of whether the DRE is found in a strict (+/- 200 bp away from a transcription start site - TSS) or loose (+/- 1500 bp away from the TSS) definition of a promoter.

To optimize model performance and prevent overfitting, I conducted an extensive hyperparameter search for each newly trained model (see Methods). Thus, for each subset of input features and for each cell line a new hyperparameter search was performed. Among all models trained during the hyperparameter search the model with the highest average performance across the five folds was selected as the representative model for the given subset of input features and given cell line. Unless otherwise stated, model performance was evaluated as the area under the Receiver Operating Characteristic (ROC) and Precision Recall (PRC) curves, averaged over five folds using the 5-fold cross validation procedure. Since the AhR binding data sets were largely

unbalanced – i.e., there was a much higher number of unbound than bound DREs (Table 2), the area under the PRC curve (auPRC) was considered as a more appropriate metric of model performance. Therefore, in each case the model producing the highest auPRC was selected as the best performing model. Nonetheless, the area under the ROC curve (auROC) was reported as it remains a useful metric to distinguish between poorly and well performing models when comparing between different cell lines or types (see Methods – Performance Metrics).

First, to investigate different input feature sets and their influence on model performance, I developed and validated models with the following feature sets used as model inputs - 1) DNase-seq only (DNase model), 2) flanking sequence only (Seq model), 3) flanking sequence and DNase-seq (Seq + DNase model), 4) flanking sequence, DNase-seq and histone modifications (Seq + DNase + HMs model), 5) flanking sequence, DNase-seq, histone modifications and transcription factor binding (referred to as the full model or Seq + DNase + HMs + TFs model). For most cell lines and types, the performance of each successive model improved, except for the primary hepatocyte (not shown) and HepG2 cells. Here, the performance of the sequence only models was overall very low, even lower than the performance of the corresponding DNase model. Nonetheless, the performance of Seq + DNase model was slightly higher than the DNase model for primary hepatocytes and HepG2. These results indicate that the flanking sequence provides some additional useful information when put in the context of the extent of chromatin accessibility (Figure 11).

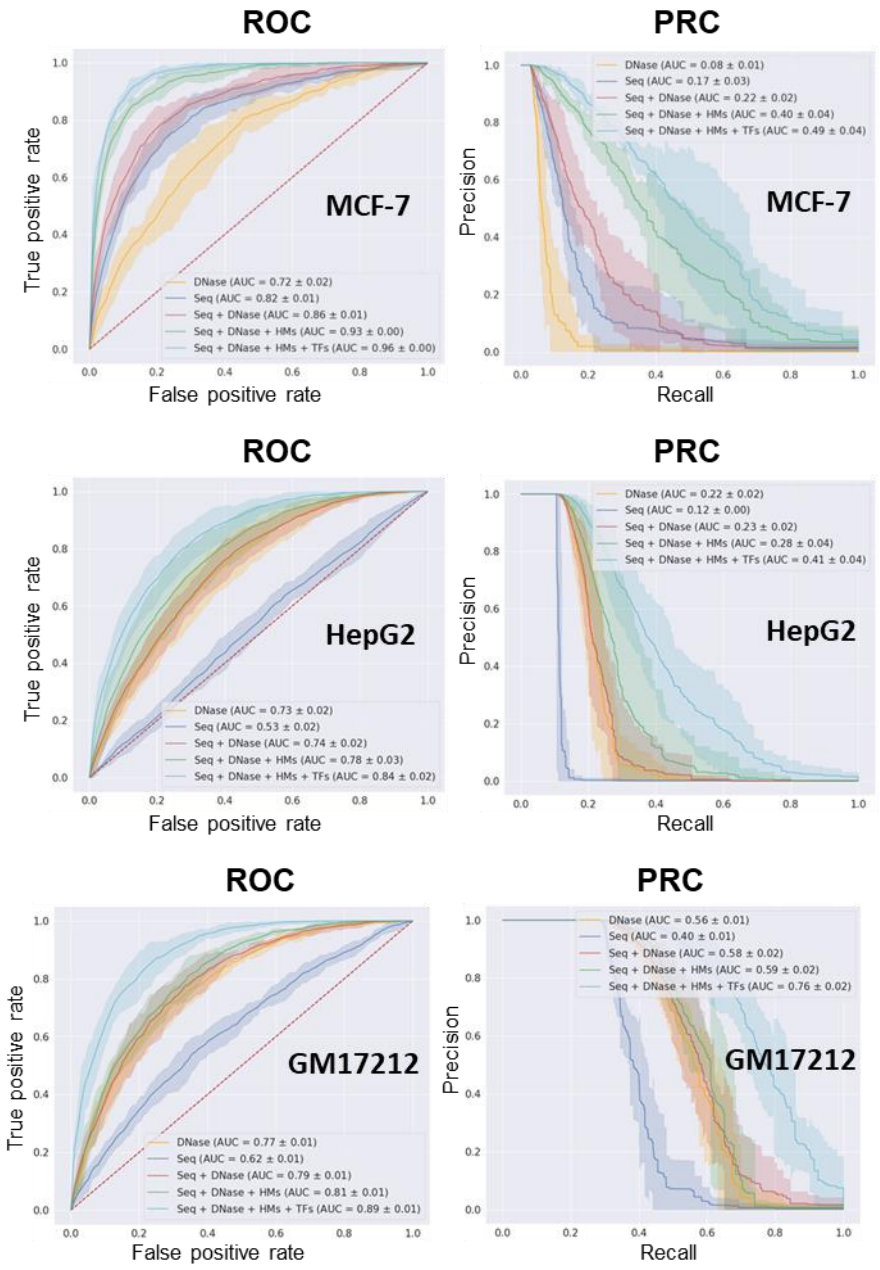

Figure 11 – ROC and PRC curves for model training.

The performance of most full (Seq + DNase + HMs + TFs) models was high, the only exception being the primary hepatocyte full model. I propose that primary hepatocyte model underperformed for two reasons. Namely, 1) lack of publicly available sequencing data in liver primary hepatocytes, and 2) the difference in the nature of hepatocytes used in different experiments. The AhR ChIP-seq experiment was performed on human primary hepatocytes

derived from a single donor specific to that experiment. All other input features, inclusive of the list of DREs in accessible chromatin, were obtained from hepatocyte-like cells in vitro differentiated from H9 cells. This discrepancy is also evidenced in a lower percentage of AhR peaks occurring in open chromatin in the primary hepatocyte experiment (Table 2).

Next, I investigated the contribution of individual chromatin context features to improving the performance of full models. For each cell line, I trained the full model on all available data with the model hyperparameters set to values previously determined to produce the best performing model. After the full models were trained, I used the information gain metric generated by XGBoost to determine the average feature importance of all features (Figure 12), the relative feature importance of sequence features per flanking sequence nucleotide position (Figure 13), and relative importance of individual bins of non-sequence features (Figure 14).

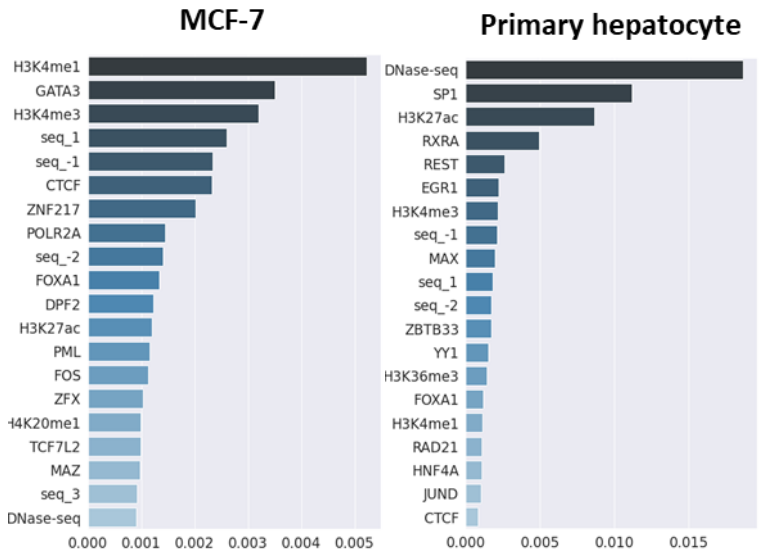

Figure 12 – feature importance lists of all features for full models.

Figure 12 (cont'd)

Figure 13 – feature importance of flanking sequence.

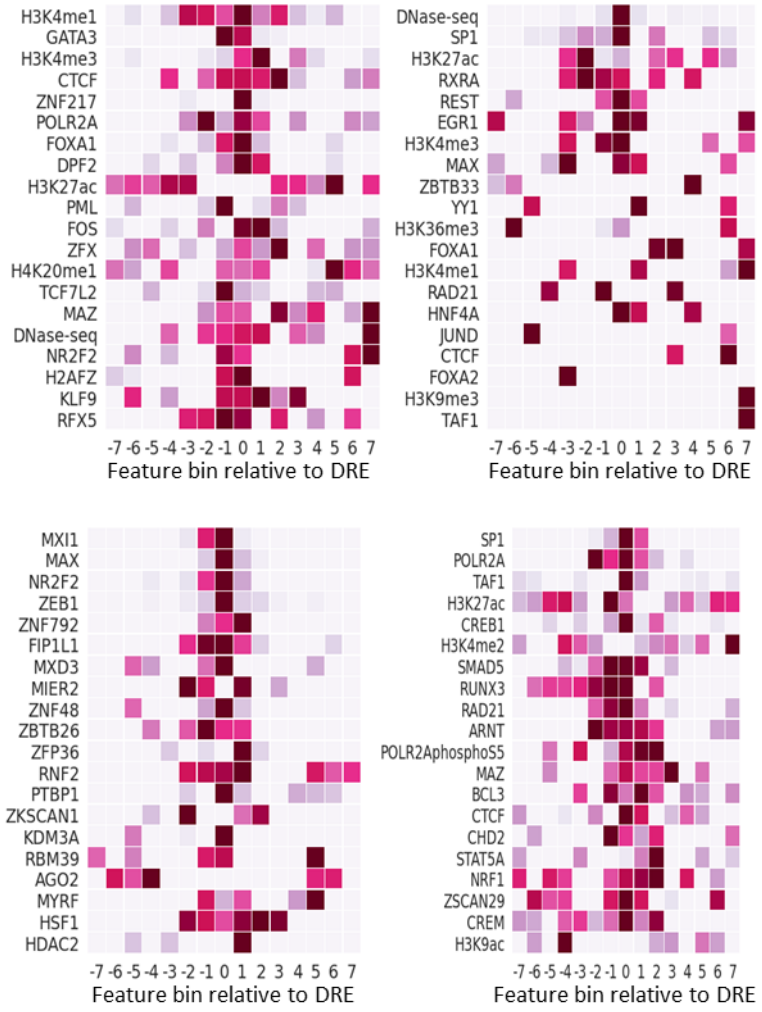

Figure 14 – feature importance of individual bins.

Examining the feature importance scores of all features used in full models I observed that specific models are predominantly learning and making AhR-DRE binding predictions by relying on different features across different cell lines and types (Figure 12). Each cell line or type had three to six bigWig signals with feature importance 2-5 times higher than that of any other signal. These were: 1) in MCF-7 cells - H3K4me1, H3K4me3, GATA3, CTCF, ZNF217 and FOXA1; 2) in primary hepatocytes - DNase-seq, SP1, H3K27ac and RXRA; 3) in HepG2 cells - MXI1, MAX, NR2F2, ZEB1; and 4) in GM17212 cells - SP1, POLR2A, TAF1, H3K27ac, and CREB1. Some features appear to be ranked relatively highly across most cell line or types, such as the binding of

CTCF, Rad21, SP1, FOXA1, MAX and MAX related factors MAZ, and MXI1; as well as histone modification H3K27ac. Nevertheless, the relative level of importance of these features varied across different cell lines or types - e.g., CTCF ranked fifth in MCF-7 cells and 20th in primary hepatocytes (Figure 12). Additionally, when looking at the feature importance scores of individual bins across cell lines or types, the distribution of relative importance scores across bins varied between cell lines or type. For example, the central bins of H3K27ac in GM17212 cells exhibited the highest importance for this feature, whereas in MCF-7 cells the central bins were not used by the model and were, consequently, not assigned an importance score (Figure 14).

To verify that the ordering of feature importance scores in Figure 12 was robust and reproducible I created ranked lists of features with highest feature importance - one ranked list for each of the five folds within the 5-fold cross validation. Next, for each feature I created a boxplot of rank distributions across the folds (Figure 15). I observe that the most highly ranked features always rank highly and thus exhibit low rank variability. For instance, H3K4me1 always ranks first in all five folds within the MCF-7 model.

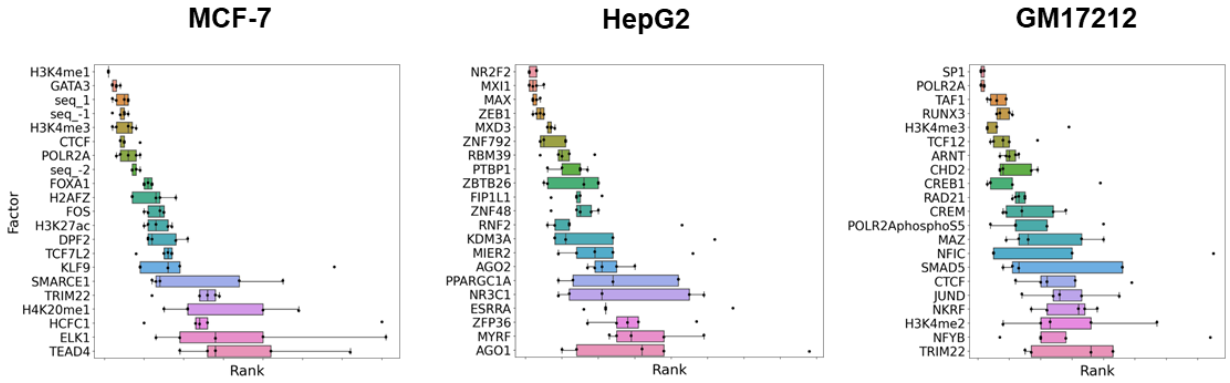

Figure 15 – ranks of features sorted by importance across five folds in 5-fold cross validation.

To investigate the contribution to model performance of DNA sequence immediately flanking the DRE, I examined the importance scores of nucleotides flanking the DRE within 1) flanking sequence-only models (not shown), and 2) full models. Importance scores of nucleotides

produced by the sequence only models were highly variable between different cell lines and types. On the other hand, the importance scores of nucleotides flanking the DRE produced by the full models demonstrated similar profiles of nucleotide importance across different cell lines or types. In summary, the thymine residue at the flanking position directly 5' of the DRE (labelled as the -1 position) had the highest feature importance out of all nucleotides that could appear at that position in three out of four examined binding experiments (Figure 13). Additionally, for two out of four binding experiments the thymine at position -2 and cytosine and guanine at position +1 also had high feature importance (Figure 13).

To examine the influence of individual TFs on model performance, I developed models that used only a single input feature – the 15 bins representing the average bigWig signal of a single TF. Sorting these models by performance, I determined that the relative ordering of transcription factors (Figure 16) was different when compared to the feature importance ranking of the full models shown in Figure 12. Notably, for MCF-7 cells (Figure 16), EP300 was the factor resulting in the second-best performing model, while in the corresponding full model, EP300 did not appear even among the top 20 features with highest importance (Figure 12). On the other hand, GATA3 was the most predictive factor both in the full model and individually (Figure 12 and Figure 16). These results point to a high likelihood of redundancy between binding of different TFs.

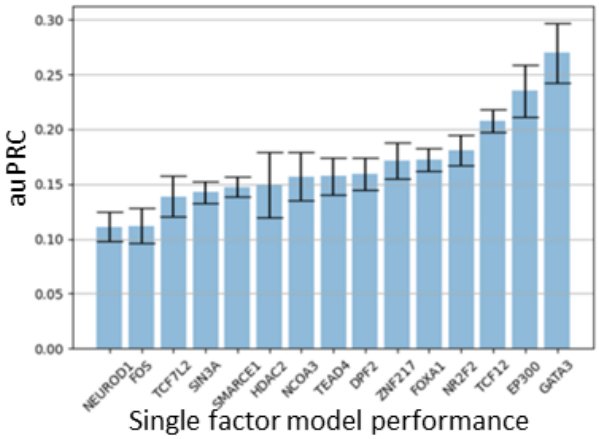

Figure 16 – single transcription factor model performance.

In conclusion, the models predicting the AhR binding status of DREs in open chromatin developed here, generalized well within the cell line or type they were trained on, when evaluated on a subset of bound singleton DREs and unbound isolated DREs left out from the training dataset (i.e., a single fold in a 5-fold cross validation). Additionally, these models exhibited highly variable chromatin context specificity between different cell lines or types. The only exceptions were the DNA flanking sequence features. Most cell lines or types exhibited similarities in DNA flanking sequence specificity, potentially pointing to a common flanking sequence grammar.

Singleton peak-trained models predict multi-DRE and 0-DRE AhR peak binding within the same cell line or cell type

To assess the robustness of trained models and to investigate the extent of overtraining, I performed feature selection based on the feature importance rankings of the full models, for each cell line or type. Briefly, using the list of 300 features with the highest feature importance in Figure 12, I created several models with increasingly larger subsets of those features used as model inputs. I created models with a subset of top N features with highest importance scores in the full model (where N = 10, 25, 50, 75, 100, 200, 300). In the MCF-7 cells I observed that the performance plateaus already at around 100 top features used and that the performance of the model using only

the top 50 features, although slightly lower on average, is not significantly different than the performance of the full model (Figure 17).

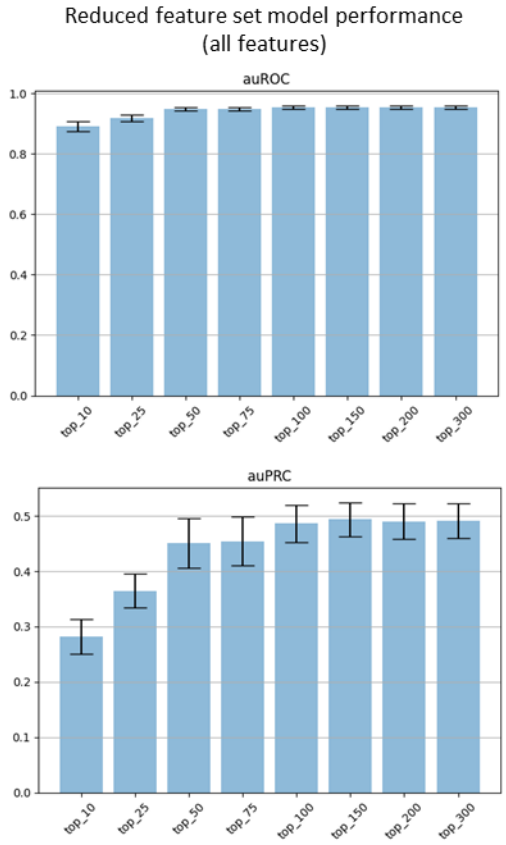

Figure 17 – reduced feature set model performance (all features).

Similarly, I investigated the influence of the number of flanking nucleotides used in model training. It was previously indicated, based on computational analysis of 13 experimentally verified DREs, that up to 7 up- and down- stream DRE-flanking nucleotides might play a role in determining AhR binding (45). Accordingly, I have created sequence only models with N flanking nucleotides up- and down- stream of the DRE used in model training (N = 1, 2, 3, 4, 5, 6, 7). The results show that the sequence-only model performance plateaus at four flanking nucleotides (Figure 18). Therefore, all models using flanking sequence features were developed using exactly four flanking nucleotides up-/down- stream of the DRE.

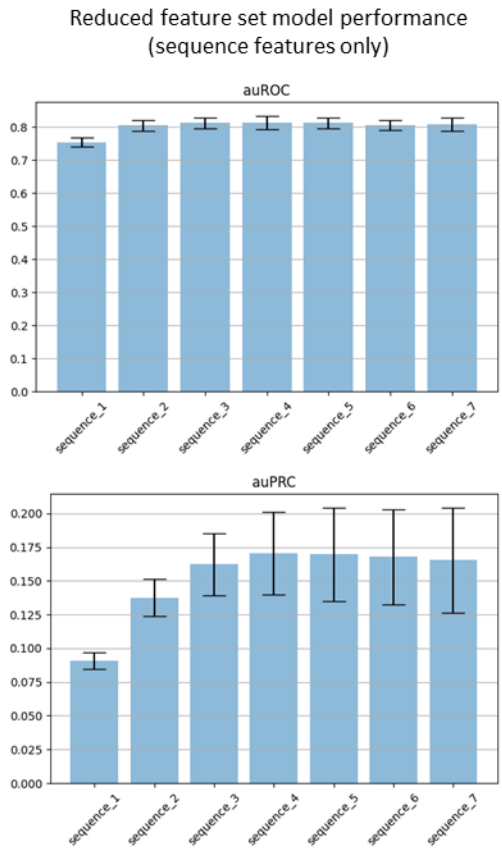

Figure 18 – reduced feature set model performance (sequence features).

Noting that the DRE is not a palindromic binding motif, it is possible that whether the DRE occurs on the forward or the reverse strand might influence AhR binding determinants. Mainly, DRE orientation might influence the spatial orientation of the AhR-ARNT heterodimer when binding DNA, and thus also influence the direction of interactions with other TFs. To account for this potential issue, I investigated whether correcting the orientation of local chromatin context features by aligning them with the orientation of the DRE influences model performance. Specifically, for DREs found on the forward strand, all features were left as-is, and for DREs found on the reverse strand, the bins of all features were flipped around the central bin (the bin containing the DRE), to match the DRE orientation. The results indicate that there are no differences between the original and strand-corrected models (Figure 19).

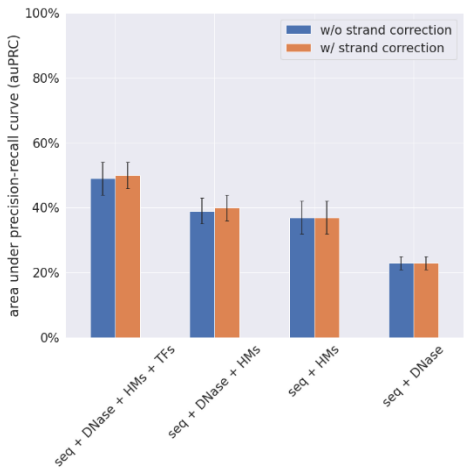

Figure 19 – model performance with and without strand correction.

Next, I evaluated the ability of full models to predict the binding status of DREs that were not used in training, as a test of how well the models generalize. Model performance was first evaluated on multi-DRE AhR peaks. For each muli-DRE peak, the binding status of each DRE was predicted by the model. If at least one DRE was predicted as bound, the peak was considered generally recovered (Figure 20 - blue bars). If the DRE closest to the summit of the peak was predicted as bound, the peak was considered centrally recovered (Figure 20 - orange bars). General recovery of multi-DRE peaks in open chromatin resulted in true positive rates (TPR) between 80-100% for most multi-DRE peaks (Figure 20).

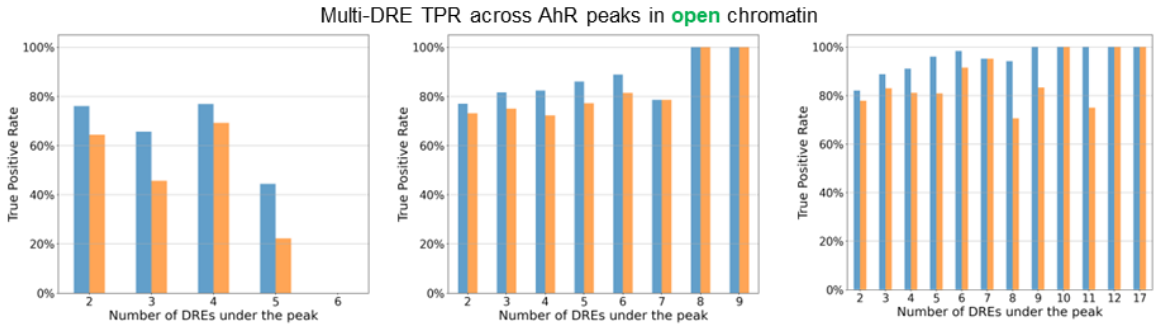

Figure 20 – model evaluation of multi-DRE AhR peaks in open chromatin.

In the MCF-7 cells, AhR peaks containing more than four DREs were mostly not recovered by the models (TPR around 40% or lower), suggesting the possibility of a different mechanism

underlying AhR binding in areas of high DRE density in MCF-7 cells, possibly through cooperative binding (95). In the hg19 human reference genome approximately 1% of all DREs can be found in one of these high DRE density areas, which were defined as 5 or more DREs within a 500-base pair region (Figure 21).

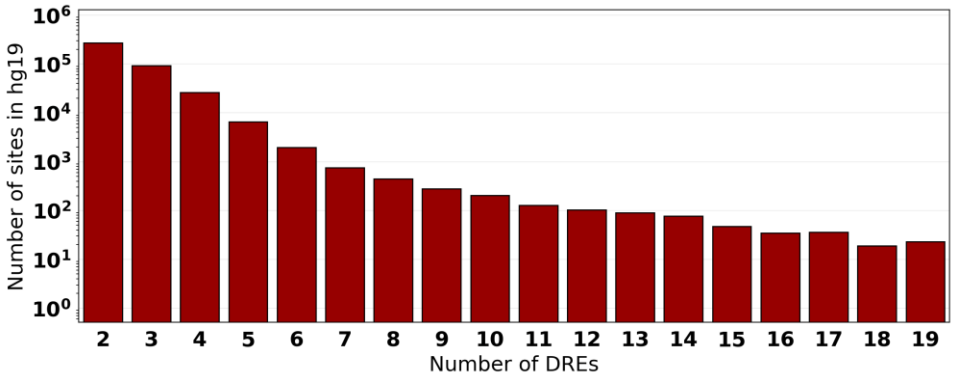

Figure 21 – number of high DRE density clusters in the human genome.

On the other hand, multi-DRE peaks in closed chromatin were recovered at a much lower and variable rate of 25-60% (Figure 22), suggesting that AhR binding in closed chromatin might be governed by a distinct set of rules compared to the binding in initially open chromatin. Alternatively, since my models are trained to predict the binding status of DREs in open chromatin, the model might struggle when predicting the binding status of DREs in closed chromatin.

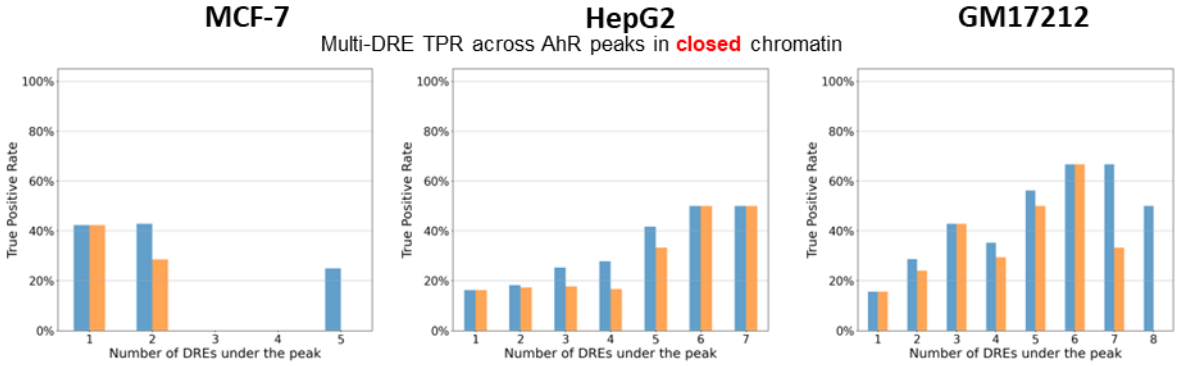

Figure 22 – model evaluation of multi-DRE AhR peaks in closed chromatin.

However, taking a closer look at DNase-seq signal in the vicinity of these closed chromatin DREs revealed no correlation between the normalized binding strength of DNase-seq signal and

the DRE binding status prediction probabilities (Figure 23). This result indicates that the lower performance of the models when predicting the binding status of bound DREs in closed chromatin was not lower due to lower DNase-seq signal alone.

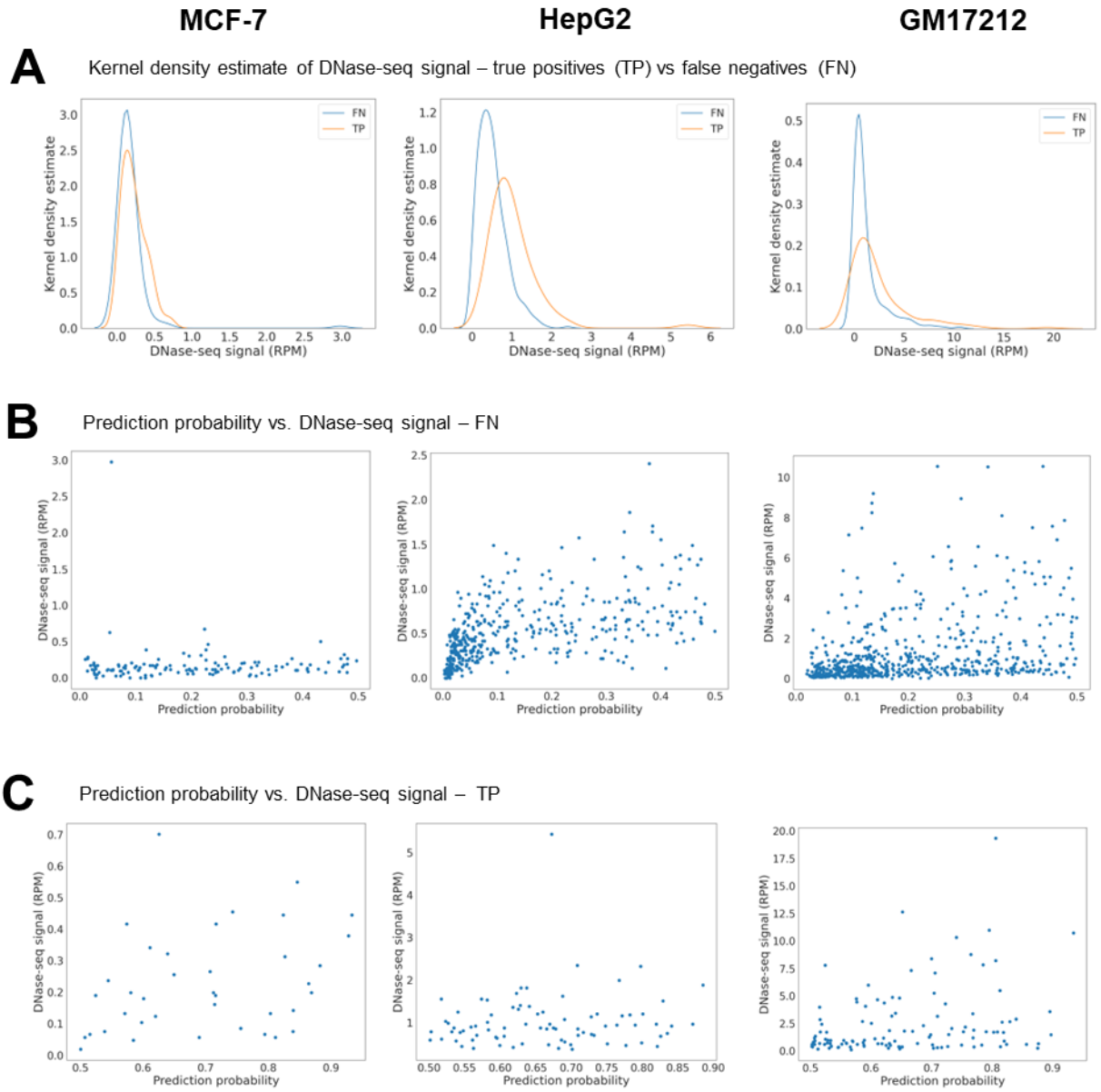

Figure 23 – DNase-seq signal vs. model prediction probabilities across multi-DRE peaks in closed chromatin.

Contrary to the expectation that the more DREs a peak contains, the higher the likelihood of that peak being recovered by pure chance, I did not observe this trend in Figure 20 and Figure

22, with either general or central recovery. These results suggest that the models generalized well, and that the multiple-testing issue was not prevalent when assessing general recovery rates.

I also evaluated the false positive rates (FPR) of the full models when predicting the binding status of 1) multi-DRE DNase-seq peaks, 2) 1+-DRE peaks of best performing TF in open chromatin, 3) 1+-DRE peaks of best performing TF in closed chromatin (Figure 24). “Best performing TF” refers to the TF that was ranked the highest in importance in Figure 12. Best performing TFs were GATA3, MXI1, and SP1 for MCF-7, HepG2 and GM17212 cells, respectively. The FPRs generally do not exceed 20% and are the lowest for multi-DRE DNase-seq peaks.

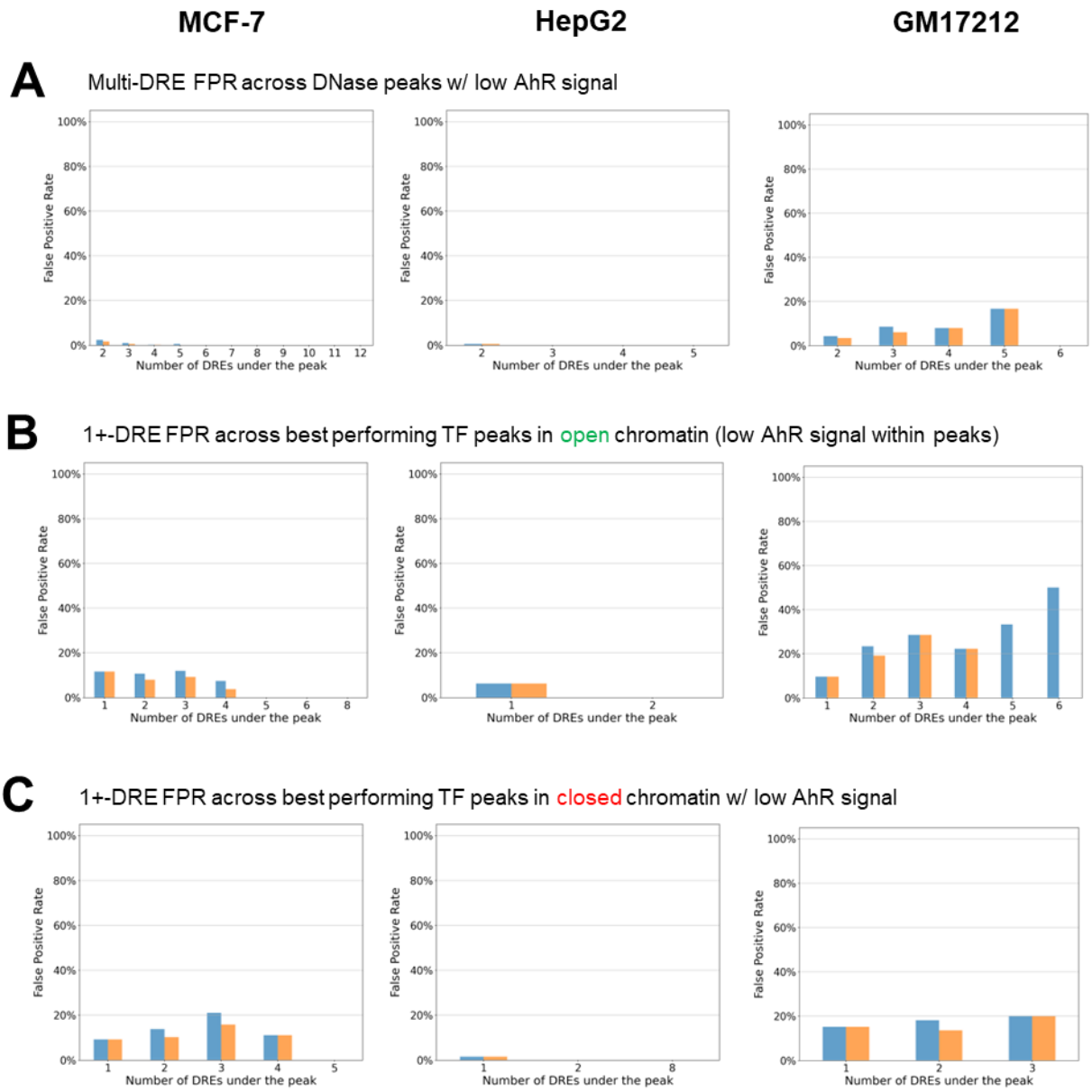

Figure 24 – model evaluation on non-AhR peaks.

Similarly, to multi-DRE peaks, I evaluated the performance of the models on 0-DRE peaks. In this case, since there were no DREs to evaluate model performance on, models trained on all features excluding flanking DNA sequence (DNase + HMs + TFs) were used for evaluation. To calculate the values of all non-sequence input features I previously used the DRE genomic location as a reference. However, since 0-DRE peaks do not have any DREs, I simulated five "dummy DREs" for each 0-DRE AhR peak to create reference points for the calculation of input features.

Unlike actual genomic DREs, these dummy DREs are not present in the genomic sequence and only define the location to be used as reference for the calculation of input features. The center of the first dummy DRE was aligned to the mid-point of the AhR peak and the other four dummy DREs were positioned at -100, -50, +50, and +100 base pairs relative to the mid-point of the peak. I investigated up to five dummy DREs for each 0-DRE peak since the majority of DREs within singleton AhR peaks were located within –100 to +100 base pairs relative to the mid-point of the peak (as shown in Figure 4). Upon establishing the dummy DREs I applied the same procedure as described for predicting multi-DRE peaks. Specifically, a 0-DRE peak was considered generally recovered if at least one of the five dummy DREs was predicted as bound. The peak was considered centrally recovered if the central dummy DRE was predicted as bound. In MCF-7 cells, approximately 93.7% of the 0-DRE peaks are partially recovered and 91% are centrally recovered. Other cell lines exhibit a slightly lower rate of recovery, nevertheless, the majority of 0-DRE peaks was recovered (Figure 25).

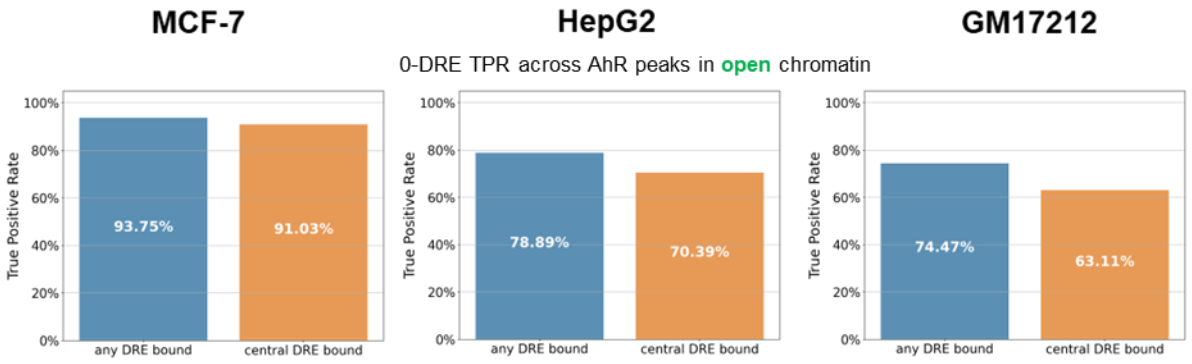

Figure 25 – model evaluation on 0-DRE AhR peaks in open chromatin.

Similar to multi-DRE peaks, the central true positive rates for 0-DRE AhR peaks in closed chromatin are considerably lower – between 11.7% and 41.8% for GM17212 and MCF-7, respectively (Figure 26).

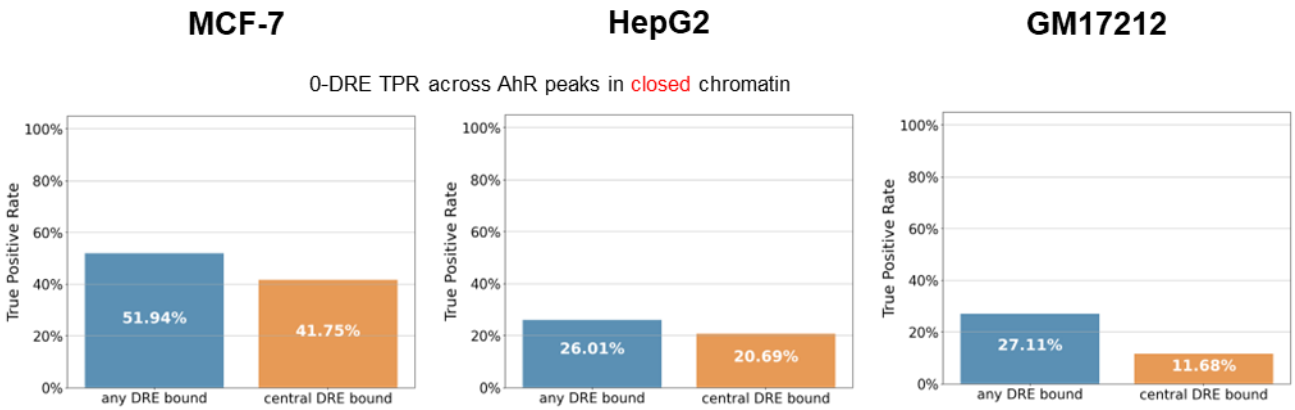

Figure 26 – model evaluation on 0-DRE AhR peaks in closed chromatin.

Additionally, I evaluated the false positive rates (FPR) of my models when predicting the binding status of 1) 0-DRE DNase-seq peaks, 2) 0-DRE peaks of best performing TF in open chromatin, 3) 0-DRE peaks of best performing TF in closed chromatin. Central FPR for 0-DRE DNase-seq peaks is relatively low and does not exceed 1.3%, However, central FPR for 0-DRE best performing TF peaks in open chromatin can be high and ranges from 8.6% to 45% for HepG2 and MCF-7 cells, respectively (Figure 27).

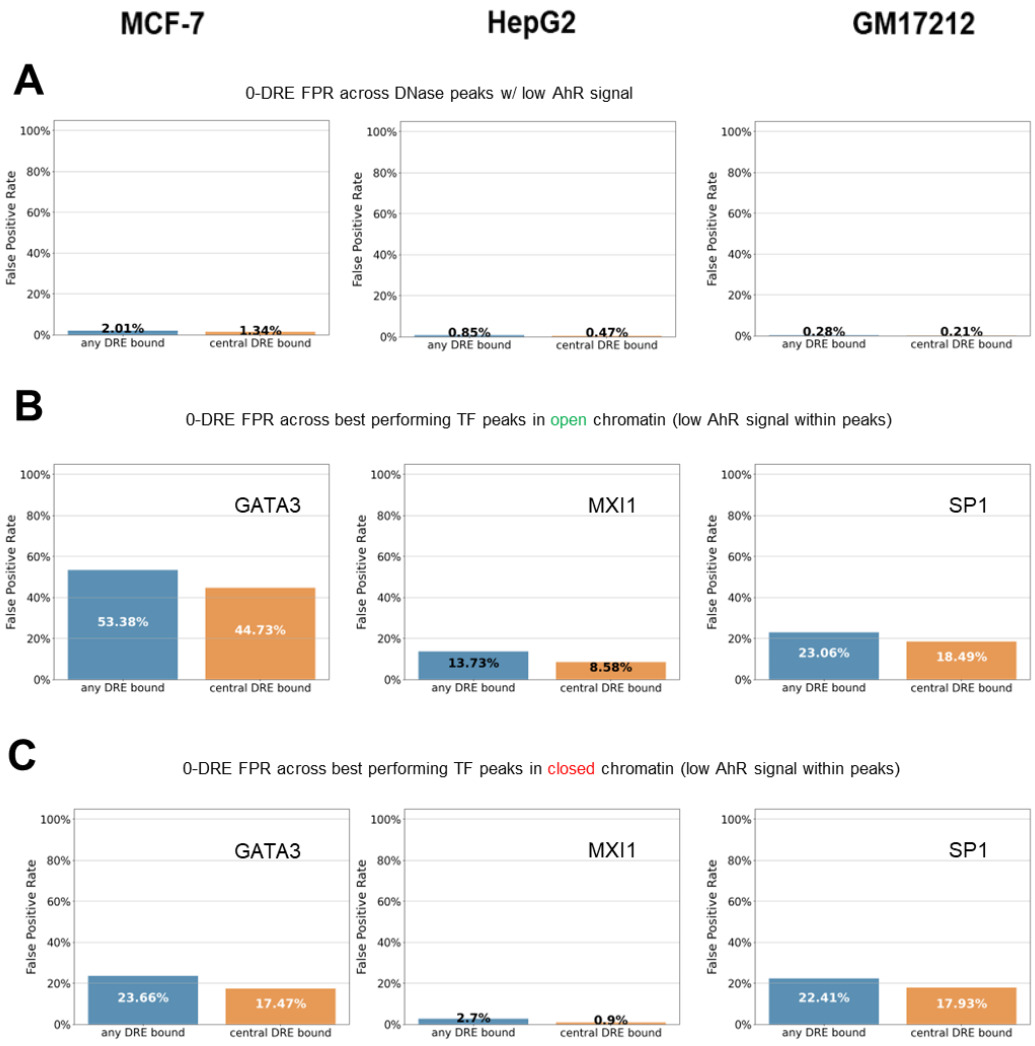

Figure 27 – model evaluation on 0-DRE non-AhR peaks.

Cross-cell models provide insights into cell-specificity of AhR binding

To examine whether full feature models for different cell lines learn from different features simply because different features were available for different cell lines, I developed full feature models that only use features available in all evaluated cell lines - MCF-7, HepG2 and GM17212. The input features were limited to DNase-seq, and to only those TF and HM features for which ChIP-seq experiments were available in all three cell lines. The results still exhibit a highly variable set of the most important features determining AhR binding within different cell lines (Figure 28).

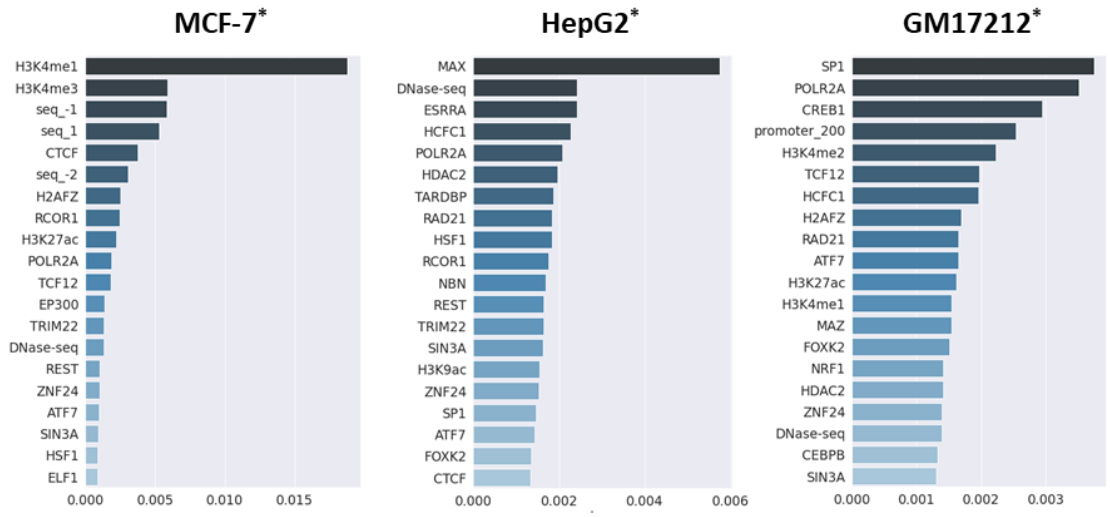

Figure 28 – feature importance lists of all features in full models for DREs in enhancers only.

On the other hand, the binding of AhR to DREs in specific genomic locations, such as promoters or enhancers, might be governed by distinct molecular mechanisms. Therefore, I investigated whether there were any enhancer-specific binding rules and whether these rules might be similar between cell lines. To this purpose, I created full models predicting the occupancy of singleton bound and isolated unbound DREs found only in enhancers. Once again, the results display a highly variable sets of the most important features for each cell line. Nevertheless, I observed an increase in feature importance for some TFs, such as the EP300 transcriptional coactivator for all cell lines (Figure 29).

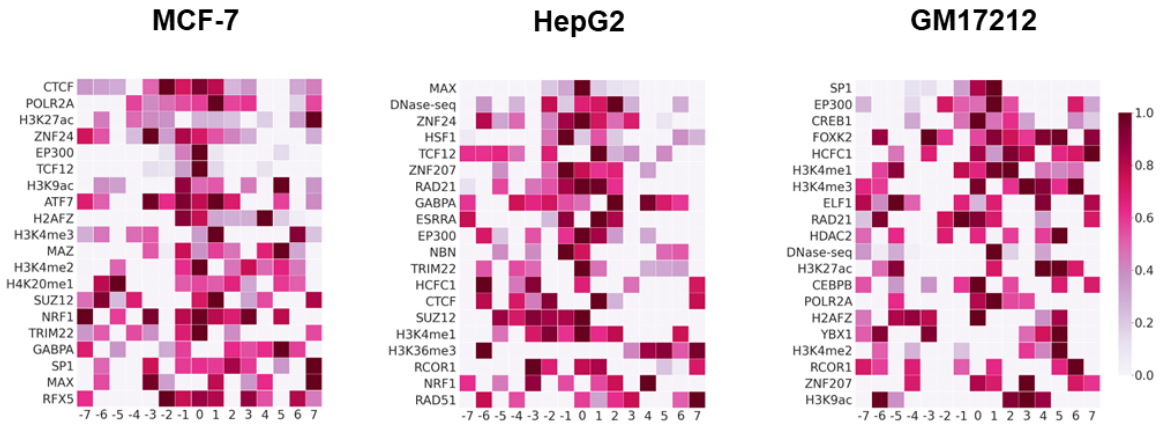

Figure 29 – feature importance lists of features shared by all three cell lines in full models for DREs.

Examining the feature importance of individual factors in different cell lines, I noticed that that SP1 and MAX transcription factor have the highest feature importance in the full models for GM17212 and HepG2 cells, respectively. However, in MCF-7 cells, neither of these factors appeared within the top 20 factors with highest feature importance (Figure 12). To investigate whether the importance of these factors might be low in MCF-7 cells due to redundancy with other TFs or other features, I examined the discriminative power of a single feature derived from these factors. For both SP1 and MAX, I created a single feature that was the maximal normalized signal within a 100-bp region surrounding the DREs used in model training. Next, for each feature in each cell line I found the optimal threshold that produced the highest F1 score. The F1 scores of both SP1 and MAX were much lower in MCF-7 cells when compared to HepG2 and GM17212 cells – 7% vs. 35% and 60%, respectively (Figure 30).

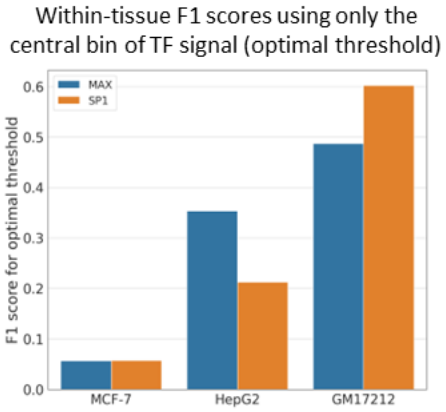

Figure 30 – within-cell line F1 scores for the optimal threshold of the central bin of the TF signal.

These results indicate that SP1 and MAX are not predictive of AhR binding in MCF-7 cells but are highly predictive of AhR binding in GM17212 and HepG2 cells. On a similar note, GATA3 is the factor most predictive of AhR binding in MCF-7 cells. Since GATA3 is not expressed in many other cell lines or types, it is difficult to investigate whether the AhR binding dependence on GATA3 in MCF-7 cells is specific to that cell line. To further investigate the difference in

predictive capabilities of features based on the MAX transcription factor I compared the binding profiles of MAX centered on bound and unbound DREs in MCF-7 and HepG2 cells. I observed that the difference in MAX signal between bound and unbound DREs appears qualitatively less pronounced in MCF-7 cells than it is in HepG2 cells, which could explain the increased utility of MAX features in the HepG2 model (Figure 31).

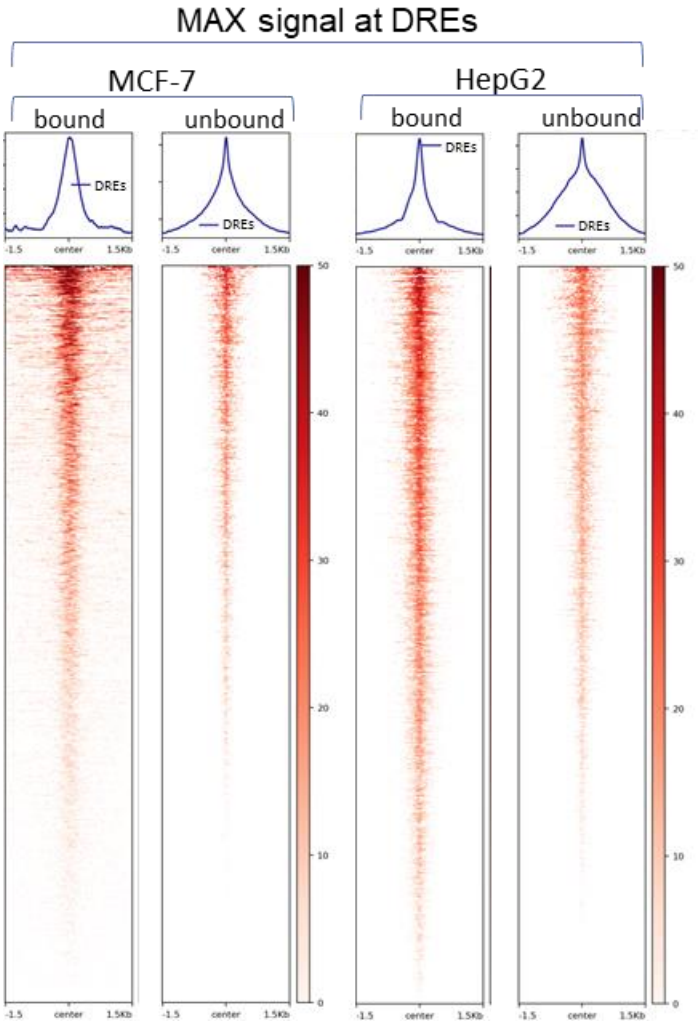

Figure 31 – heatmaps representing MAX binding across bound and unbound DREs in MCF-7 and HepG2 cells.

Next, I evaluated the cross-cell performance of two models with similar treatments – MCF-7 and primary hepatocyte models. Here I focused on the sequence-only models to examine the possibility of a cross-cell flanking sequence grammar. The sequence-only model trained on MCF-

7 cells and evaluated on primary hepatocytes did not perform any better than random. Conversely, the sequence-only model trained on primary hepatocytes performed better when evaluated cross cells on MCF-7 cells than within-cells in primary hepatocytes (Figure 32). Admittedly, the cells in these two experiments were both treated with TCDD for 24 hours, although the concentration of TCDD was different - 10 nM and 1 nM, for MCF-7 cells and primary hepatocytes, respectively.

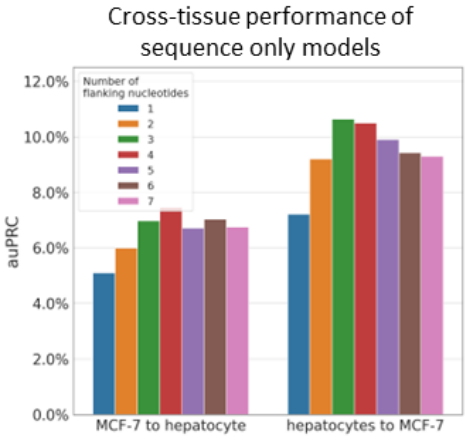

Figure 32 – cross-cell performance of sequence only models.

To further evaluate cross-cell performance, I focused on two similar cell lines – MCF-7 and T-47D, which are both epithelial breast cancer luminal type A cell lines. Models comprising of 1) sequence-only, 2) GATA3-only, and 3) sequence and GATA3 features trained in the MCF-7 cells exhibit high performance within cells (results not shown), where model performance rises with each successive model. Nonetheless, when these models were evaluated on all bound DREs within T-47D cells treated with either 3-MC or TCDD for one hour, I found that it was the sequence-only model that had the highest true positive rate of 90.67%. Any DREs that were bound in both MCF-7 and T-47D cells were excluded from model evaluation to avoid using training data in the testing phase. When evaluating the sequence-only model trained on MCF-7 cells separately in three subgroups of T-47D DREs, namely DREs bound in 1) TCDD treatment only, 2) 3-MC treatment only, 3) and both TCDD and 3-MC treatment, the true positive rates were 90.48%,

89.47%, and 92.86%, respectively (Figure 33). Therefore, the sequence-only model could not distinguish between AhR binding resulting from activation by different AhR ligands. These results indicate that AhR binding resulting from activation by different ligands in the same cell line might not be as different as Figure 7 seems to suggest. It is therefore possible that many of the AhR peaks unique to a single ligand were also sub-threshold peaks for the other ligand.

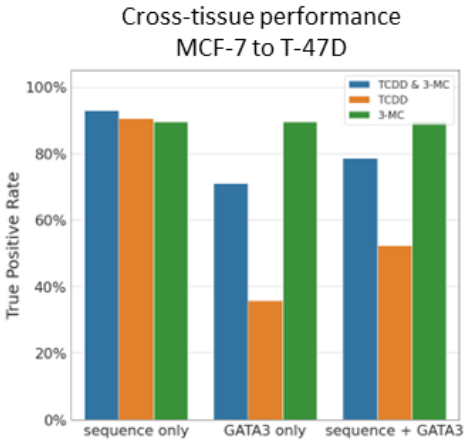

Figure 33 – cross-cell performance of MCF-7 models applied on T-47D cells.

Taken together, these results also suggest that cross-cell model predictions might be more accurate for binding experiments with similar cell lines or types (like MCF-7 and T-47D, both of which are breast carcinoma cell lines) and treatments, as opposed to binding experiments with dissimilar cell types (like MCF-7 and primary hepatocytes) but similar treatments.

AhR binding models reveal positive and negative regulators of AhR binding

Next, I focused on analyzing individual DRE binding status predictions. To this end, I used ELI5 - https://eli5.readthedocs.io/, an algorithm that summarizes the decision-making process underlying individual model predictions. ELI5 assigns a numerical weight to each feature the model used when making each DRE binding status prediction. These feature weights are a summary measure of how much the feature contributed to the final DRE binding status prediction across all decision trees used by the XGBoost model. The higher the weight the more the feature

contributed. Features can be both positively and negatively weighted and an example showcasing the top 10 positively and negatively weighted features for a single DRE binding status prediction is shown in Figure 34. In this example, the high average bigWig signal value within bin 0 of MAX binding is assigned the highest weight by ELI5 which means that this feature contributes the most to the model predicting the corresponding DRE as bound.

Figure 34 – an example showcasing the top 10 positively and negatively weighted features for a single DRE.

When analyzing the weights assigned to features that are the individual bins of bigWig signals across all correctly predicted bound DREs, i.e., true positives (TPs), I observed both model features whose weights increase with increasing feature values - termed positive regulators, as well as features whose weights decrease with increasing feature values – termed negative regulators (Figure 35).

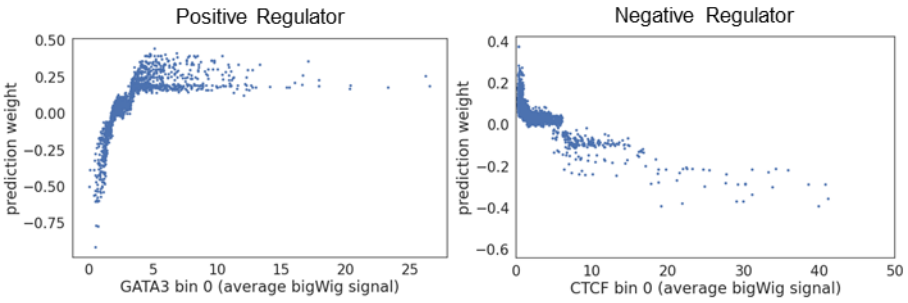

Figure 35 – scatterplot examples of positive and negative regulators.

Upon determining the direction of regulation – positive or negative regulator, for each feature in each cell line, I compared the direction of regulation of features in common to each combination of two cell lines. In total I found 102, 46, and 59 features used in common by 1) MCF-7 and HepG2, 2) MCF-7 and GM17212, and 3) HepG2 and GM17212 models, respectively. I observe that for each model combination approximately 41-62% of features appear as positive regulators in one cell line and negative in another (Figure 36).

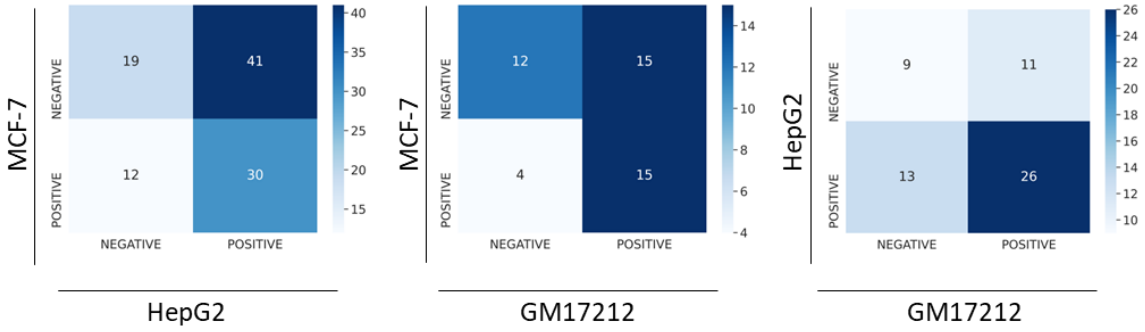

Figure 36 – positive and negative regulators in different pairs of cell lines.

Further, a total of 15 features was used by all three models. Out of those, only three features had the same direction of regulation in all three models – bin 1 of MAZ, bin 1 of MAX, bin 3 of H3K27ac. All three of these features are positive regulators in all three models (results not shown). Even different bins of a single transcription factor, e.g., CTCF in HepG2 cells, can be both positive and negative regulators, albeit the bins that are negative regulator had very small feature weights (Figure 37). These results suggest that even though cell-specific models primarily learn from entirely different features, a small subset of those features shows similar patterns across cells. Additionally, different TFs might both facilitate or interfere with AhR binding in different cell lines.

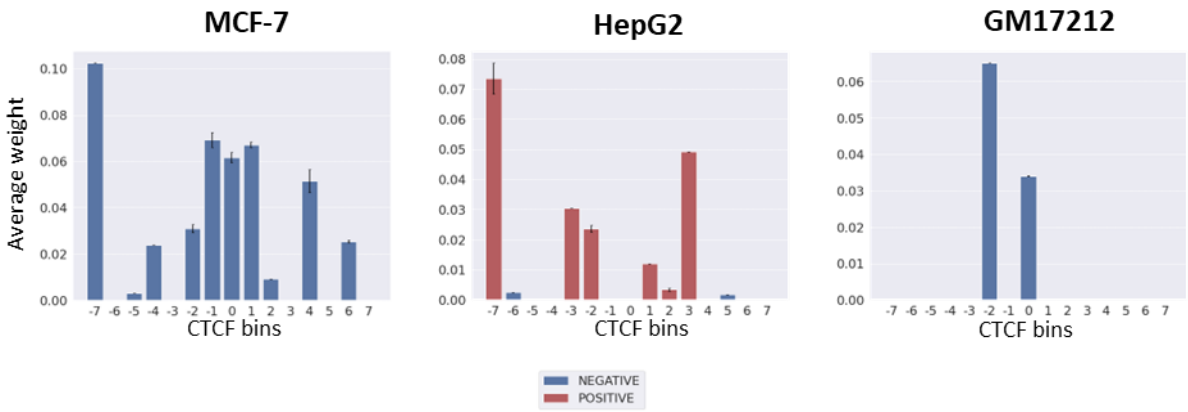

Figure 37 – average weights assigned to different bins of CTCF.

Lastly, ELI5 provided weights for flanking sequence features, as well. The flanking sequence feature were binary – i.e., a certain type of nucleotide either appeared at a specific position or it did not, e.g., nucleotide at position -1 was a thymine or not. The flanking sequence features can also be seen as positive or negative regulators, as the model produces either positive or negative weights when a specific nucleotide at a particular position is present. Such sequence features are classically represented in the form of a sequence logo (95). These representations indicate how informative the presence of a certain nucleotide at a particular position is when determining whether the given sequence is a binding site or not. However, when these logos are formed, usually only the bound sequences are considered. Conversely, the logo generated by my models is a combination of two motifs – one that describes bound DREs, and another that describes the unbound DREs (the upper and lower motifs in Figure 38, respectively). I propose that this type of motif is more informative that a standard TF binding motif.

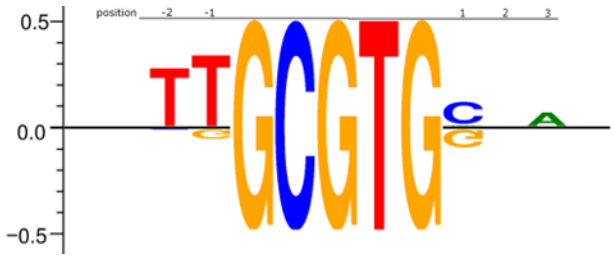

Figure 38 – motif logo representing bound and unbound DREs.

Unlike non-sequence features that are common to different cell line models, the direction of regulation for DNA flanking sequence features appears more stable across cell lines or types. For instance, between the MCF-7 cells and primary hepatocytes, all three nucleotide features (position -2 is thymine, -1 is thymine and 1 is guanine) that are used by both models have the same direction of regulation (results not shown).

CHAPTER 3: AHR BINDING PARTNERS

INTRODUCTION

Some transcription factors (TFs), such as the AhR, are incapable of binding DNA by themselves due to their incomplete DNA binding domain, and need to dimerize with other TFs to bind DNA. The primary and most widely investigated dimerization partner of AhR is the AhR nuclear translocator (ARNT) protein. The AhR-ARNT dimer tends to dominate the AhR-DNA binding landscape – hence ARNT has been regarded as the canonical dimerization partner of AhR (96). In addition, there is little evidence of in vivo AhR binding in the absence of ARNT. However, many dimerizing TFs have multiple possible dimerization partners, for instance ARNT can dimerize with itself (97), AhR (96) and HIF1α (98). It is thus possible that AhR also has multiple dimerization partners and could potentially even bind to different cognate sequences when dimerized with different partners.

Recently, it has been shown that AhR binds certain loci in the absence of ARNT, even when treated with exogenous ligands, such as 2, 3, 7, 8 tetrachlorodibenzo-*p*-dioxin (TCDD). One such locus exists in the promoter of the plasminogen activator inhibitor 1 (PAI-1) gene. The AhR binds this locus in a TCDD-inducible manner, however the binding of ARNT is markedly absent. In addition, the promoter of PAI-1 possesses no dioxin response elements (DREs) – the 5'-GCGTG-3' core consensus binding motif of the AhR-ARNT dimer (99). Huang and Elferink investigated preserved sequences across species in the promoter of PAI-1 and identified two likely locations for the binding of AhR. By mutating these sequences and testing for binding via electrophoretic mobility shift assay (EMSA) they narrowed down AhR binding to a single region and identified several nucleotides within that region that influence AhR binding. This region was termed the nonconsensus DRE (NC-DRE) and it shared marked homology with the DNA binding

sequence of the Krüppel-like factor (KLF) family of TFs. Later, it was confirmed that KLF6 interacts with AhR and binds to the NC-DRE in the PAI-1 promoter in a TCDD-dependent manner. Furthermore, sequential deletion studies demonstrated that the C terminus of the AhR and the N-terminal domains of KLF6 are necessary to facilitate this interaction (29).

The activities of AhR and NF-kB pathways have also been functionally linked (100). Tian et al. demonstrated that AhR and the RelA subunit of NF-kB associate physically in murine hepatoma cells. Additionally, such physical interactions between AhR and RelA in the absence of ARNT have also been shown in the IL-6 promoter of human lung cells (28), and c-myc promoter in breast cancer cells (27).

However, both KLF6 and RelA interactions with AhR have only been demonstrated at a limited number of loci and it is currently unknown if they could be more widespread. To investigate this possibility, I have developed a computational method to assess the likelihood of AhR interactions across the entire genome, by using publicly available ChIP-seq data. My results indicate that while TCDD-activated AhR predominantly interacts with ARNT, at a subset of sites, TCDD-activated AhR appears to bind with RelA as well. On the other hand, in cells not explicitly treated with an AhR ligand, AhR does not seem to interact with ARNT, except for a small subset of AhR peaks with DREs (0.5% of all AhR peaks). In this case, the AhR does seem to interact with both KLF6 and RelA extensively, across the genome.

MATERIALS AND METHODS

Reference genome

The reference genome used for sequence alignment in this part of the thesis was the human genome assembly version hg38.

Genomic locations of DREs

A list of all DREs and their genomic locations in the human genome was compiled by searching the hg38 human reference genome sequence for all occurrences of the core DRE sequence 5'-GCGTG-3' on either strand of the DNA.

Scatterplot of signal-to-signal correlation between two TFs

Given 1) a list of genomic ranges, e.g., a list of TF binding peaks, and 2) two TFs - TF1 and TF2, and their bigWig binding strength signal files representing the genome-wide intensity of TF-DNA binding; a scatterplot of binding strength signal correlation is constructed in the following way. For each genomic range, described by the chromosome, start and end of the range, the maximum of the binding strength signal of both TF1 and TF2 is found. A point is plotted on a scatterplot, where the x-axis represents the binding strength signal of TF1 and the y-axis represents the binding strength signal of TF2. The procedure is illustrated below (Figure 39).

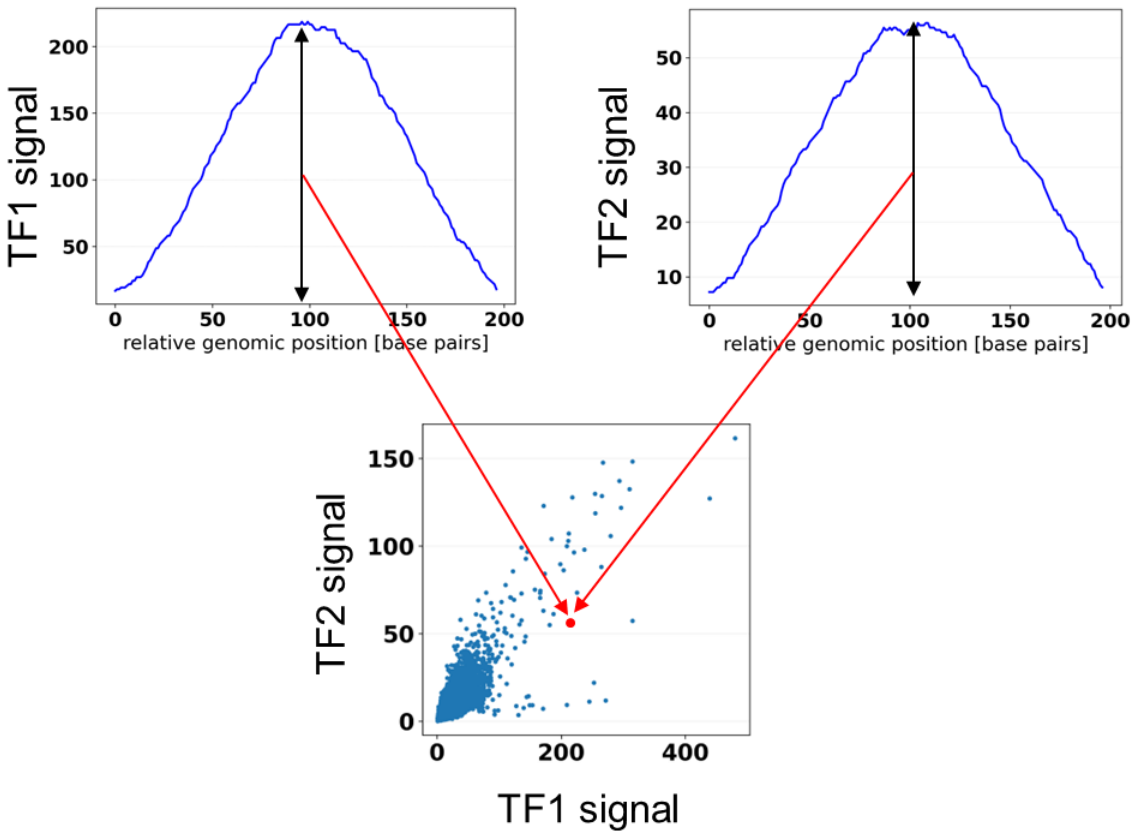

Figure 39 – construction of the scatterplot of signal-to-signal correlation between two TFs.

Histogram of individual signal correlations between two TFs

Given 1) a list of genomic ranges, e.g., a list of TF binding peaks, and 2) two TFs - TF1 and TF2, and their bigWig binding strength signal files representing the genome-wide intensity of TF-DNA binding; a histogram of signal-to-signal correlations is constructed in the following way. For each genomic range, described by the chromosome, start and end of the range, the binding strength signals of both TF1 and TF2 within the given genomic range are extracted and converted into a numerical series of values. The Pearson correlation coefficient for the two series is calculated and recorded. A histogram of all Pearson correlation coefficients for all genomic regions of interest is constructed. The procedure is illustrated below (Figure 40).

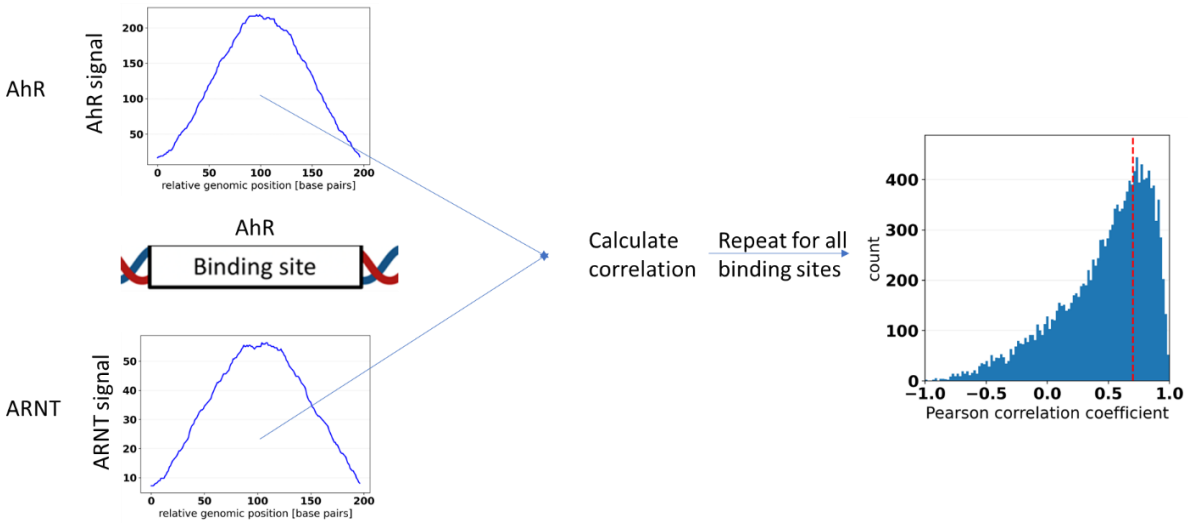

Figure 40 – construction of the histogram of individual signal correlations between two TFs.

RESULTS

Genome-wide investigation of protein-protein interactions for DNA-bound TFs

Certain TFs, like the AhR, are generally considered incapable of binding DNA by themselves and need to dimerize with other proteins to do so. Many such factors have more than one possible dimerization partner. For instance, ARNT can dimerize itself (97), AhR (96) and HIF1α (98) and possibly other TFs as well. Two TFs that dimerize to bind DNA, e.g., TFs A and B, do so together and could, therefore, be considered a single new A-B TF that is bound to DNA. The dimerization reaction, as well as dimer-DNA binding, are reversible reactions, however the crosslinking procedure that is the first step of the chromatin immunoprecipitation (ChIP) type of experiments makes the dimer-DNA complex stable. Thus, two ChIP-seq experiments, one for TF A and another for TF B should appear as though they were two replicate ChIP-seq experiments for the same A-B TF. Namely, they would appear as two biological replicate experiments that were also performed with different antibodies (Figure 41). However, this would be true only within the context of DNA sites that were bound by the dimer and not by individual TFs, or by an individual TF dimerized with a different TF.

To investigate TF-TF interactions across TF-DNA bound sites I propose an analytical method based on two semi-qualitative metrics. These are 1) the scatterplot of signal-to-signal correlation and 2) the histogram of individual signal correlations. Both metrics are constructed starting from a list of peaks. This list could be (i) a list of peaks of TF A, (ii) list of peaks of TF B, or (iii) the intersection of list of peaks of TF A and TF B.

Given the list of peaks, the *scatterplot of signal-to-signal correlation* is generated by calculating the maximum of TF A and TF B signal for each peak and plotting these two maximums as a point on the scatterplot (left panels in TF-TF interaction figures). Each point represents a single peak (Figure 39). The value of the Pearson correlation coefficient – r, for all points on the scatterplot was also reported (see Methods for more details). Similarly, given a list of peaks, the *histogram of individual signal correlations* is generated by calculating the Pearson correlation between the signal of TF A and TF B across each peak, by first transforming these signals into number series of equal length, and then calculating their correlation coefficient. All Pearson correlation coefficients are then plotted on a histogram (Figure 40). The percentage of peaks having Pearson correlation coefficient $r>0.7$, was reported on the graph as well (see Methods for more details).

This analytical method possesses an advantage over experimental methods such as immunoprecipitation followed by mass spectrometry (IP-MS) which are used to investigate protein-protein interactions. The IP-MS method pulls down the protein of interest (POI) and then performs mass spectrometry to obtain a list of proteins interacting with the POI. However, this method assumes that the two proteins interact even when not bound to the DNA, which might not be the case. The proposed method should work even if the concentration of the dimerized protein was generally much lower than the concentration of individual proteins A and B across the cell.

This is because the method works by selecting for sites bound by the A-B dimer, enriching for the A-B dimer signal in the process.

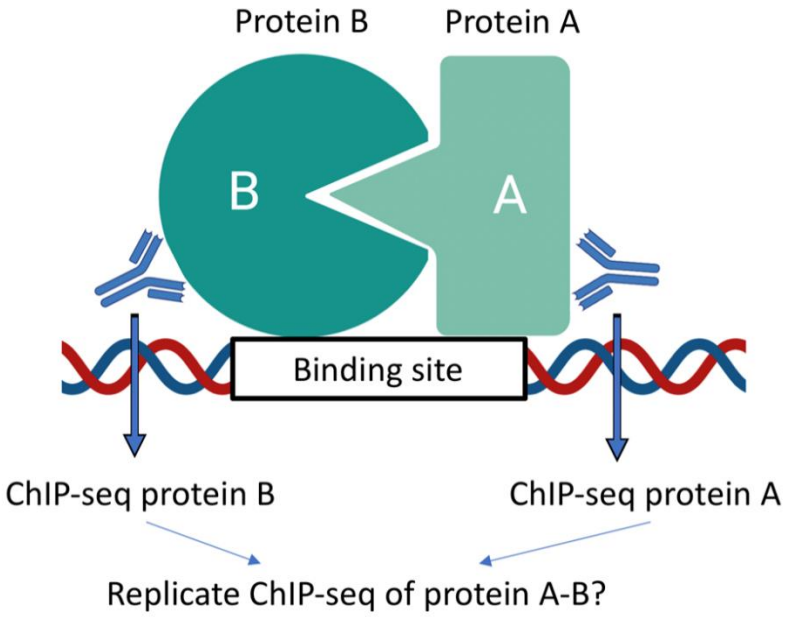

Figure 41 – ChIP-seq of dimerized TFs as replicate ChIP-seq of a single TF.

To set some expectations for these metrics in different scenarios, I first generated the scatterplot and histogram for two replicate experiments for the same TF – FOXA1. Data was obtained from ENCODE. The results demonstrated that for replicate experiments one could expect the scatterplot correlation to be high, r=0.82 in case of FOXA1 replicate experiments (left panel Figure 42). Similarly, the proportion of peaks with signal correlations exceeding 70% (r>0.7) was very high – 70% in the case of FOXA1 replicate experiments, and considerably shifted to the right (right panel Figure 42).

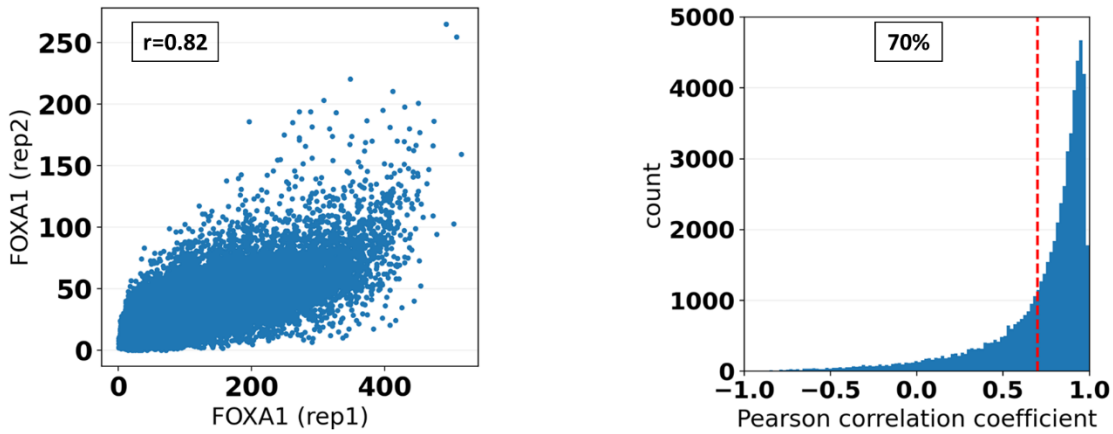

Figure 42 – FOXA1 replicate experiment interaction assessment – example of interacting TFs.

Then, I generated the scatterplot and histogram for two experiments of TFs that are known not to interact – FOXA1 and CTCF. Data for both obtained from ENCODE. The results demonstrate that for experiments with non-interacting factors one could expect the scatterplot correlation to be very low, r=0.08 in this case (left panel Figure 43). Similarly, the proportion of peaks with signal correlations exceeding 70% (r>0.7) was very low – 9% in this case, and the histogram is relatively flat (right panel Figure 43).

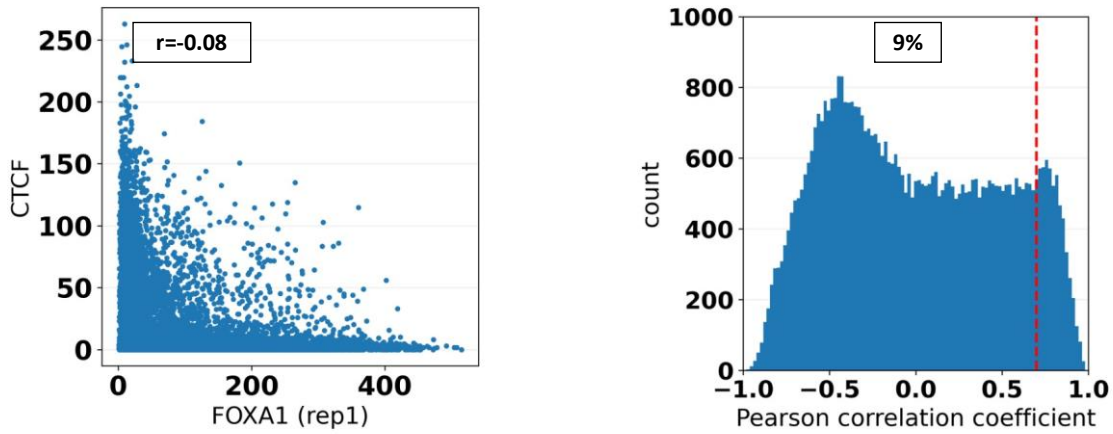

Figure 43 – FOXA1 and CTCF interaction assessment – example of non-interacting TFs.

AhR interactions with ARNT

To confirm and further investigate AhR interactions with ARNT, I examined two pairs of AhR and ARNT experiments (Table 3). The first pair of AhR and ARNT experiments was

conducted in MCF-7 cells treated with 10 nM TCDD for 24 hours under the same conditions by the same lab. The second pair of experiments assessed AhR and ARNT binding in HepG2 cells not treated by an AhR ligand (data available on ENCODE portal).

Cell line	AhR	ARNT
MCF-7	10 nM TCDD for 45 minutes	10 nM TCDD for 45 minutes
HepG2	No treatment	No treatment

Table 3 – AhR and ARNT ChIP-seq experiment list.

AhR-ARNT interactions in TCDD-treated MCF-7 cells across all AhR peaks. In scenarios of treatment with exogenous AhR ligands such as TCDD, the AhR is assumed to require dimerization with ARNT to interact with DNA. On the other hand, ARNT is known to bind DNA by dimerizing with other TF partners, such as with itself or Hif1a. Therefore, I focused on investigating AhR-ARNT interactions by using the list of 17,588 AhR peaks in MCF-7 cells treated with 10 nM of TCDD for 45 minutes. I observed that the scatterplot correlation was very high, r=0.88 (left panel Figure 44), comparable to results for the two replicate experiments. The proportion of peaks with signal correlations exceeding 70% (r>0.7) was 26%, and the histogram was slightly shifted to the right (right panel Figure 44). This result is higher than for non-interacting TFs, but lower than for two replicate experiments.

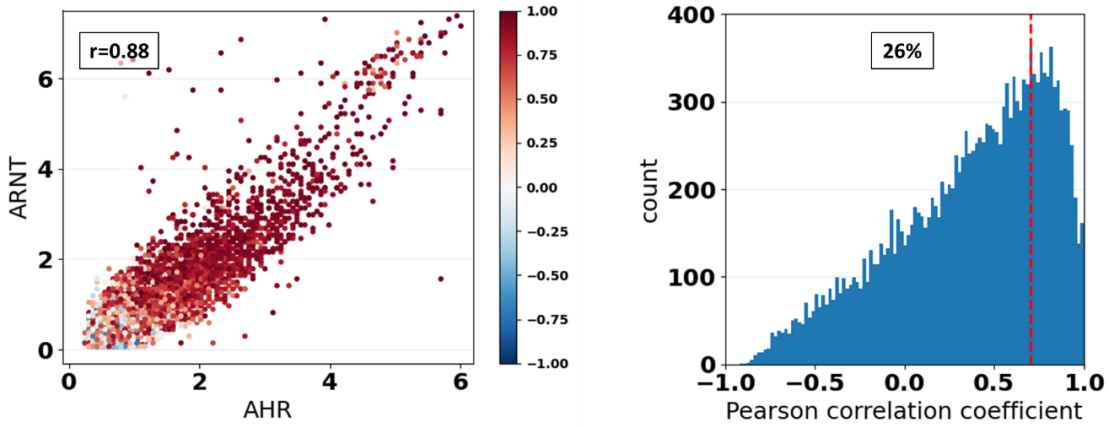

Figure 44 – AhR-ARNT interactions in TCDD-treated MCF-7 cells across all AhR peaks.

AhR-ARNT interactions in TCDD-treated MCF-7 cells across AhR peaks with 1+ and 2+ DREs. Next, I investigated how the AhR binding motif, also known as the dioxin response element (DRE), influences the AhR-ARNT interaction results. I have again generated the scatterplots and histograms of TF interactions, but this time for **(1)** AhR peaks with 1 or more DREs – 1+ DREs, 3097 AhR peaks (left panel Figure 45), and for **(2)** AhR peaks with 2 or more DREs – 2+ DREs, 563 AhR peaks (right panel Figure 45). The results indicate that the scatterplot correlation increases with increasing number of DREs, r=0.90 and r=0.93 (top left and right panels Figure 45). I also observed the narrowing of the scatterplot, with less variation with increasing number of DREs under AhR peaks. The proportion of peaks with signal correlations exceeding 70% (r>0.7) also increased to 36% and 43%, for 1+ DREs and 2+ DREs AhR peaks, respectively (bottom left and right panels Figure 45).

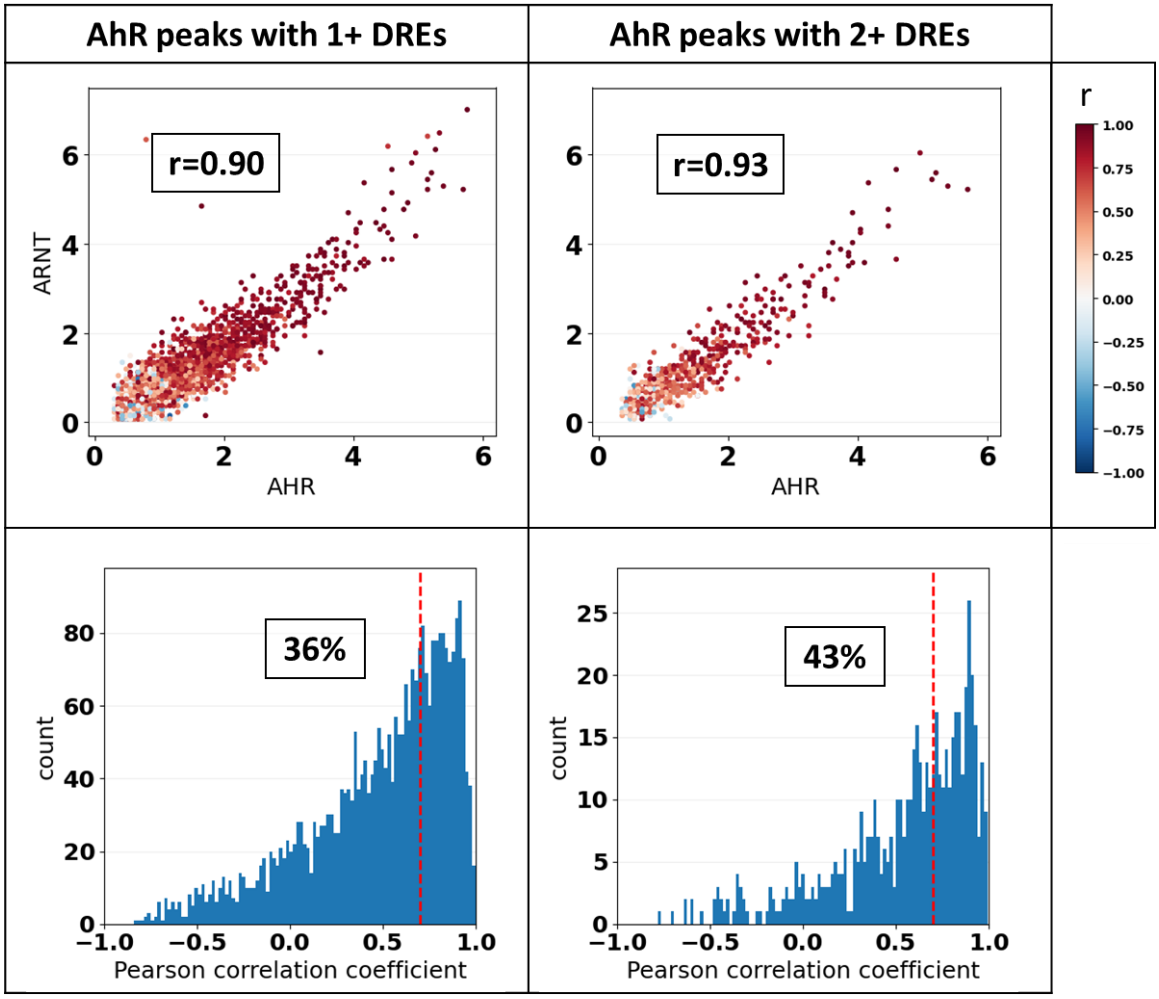

Figure 45 – AhR-ARNT interactions in TCDD-treated MCF-7 cells across AhR peaks with 1+ and 2+ DREs.

AhR-ARNT interactions in TCDD-treated MCF-7 cells across ARNT peaks that do not overlap AhR peaks. Further, I investigated the correlations between AhR and ARNT binding across ARNT peaks that do not overlap AhR peaks – ARNT-only peaks. As mentioned, ARNT readily dimerizes and binds DNA with TFs other than AhR, hence I expected to see lower degrees of correlation between AhR and ARNT binding. The scatterplot correlation decreased to r=0.63 (left panel Figure 46) compared to AhR peaks which was r=0.88. A trend similar to AhR peaks is still observable, however this is likely due to some AhR peaks not being called by the peak caller, despite possessing high AhR signal. On the other hand, the proportion of peaks with signal

correlations exceeding 70% (r>0.7) is low at 12%, making it more comparable to non-interacting factors (right panel Figure 46).

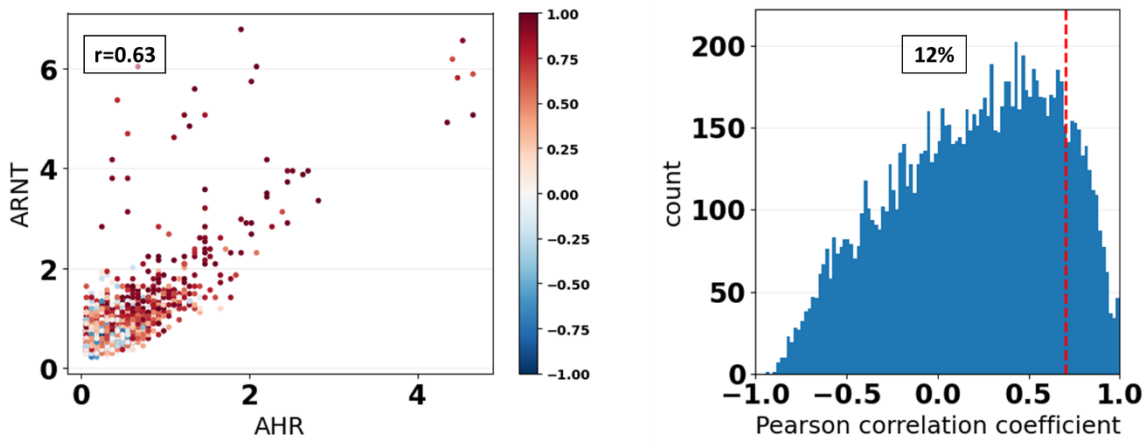

Figure 46 – AhR-ARNT interactions in TCDD-treated MCF-7 cells across ARNT-only peaks.

To investigate how AhR signal influences these results I have split the ARNT-only peaks into two groups. The first group is referred to as the low AhR signal group – with AhR signal lower than the AhR peak with the lowest AhR signal. The second group is referred to as the high AhR signal group – with AhR signal higher than the AhR peak with the lowest AhR signal. I observed that about 13% of ARNT peaks had no AhR signal at all, and that 61% fell into the low AhR signal group. Therefore, the left panel in Figure 47 contains more peaks than the right panel, even though that might not be obvious. The correlation coefficient of the signal scatterplot was much lower in the low AhR signal group than the high AhR signal group, r=0.11 vs. r=0.72, respectively (Figure 47). In addition, none of the ARNT peaks with low AhR signal had more than 2 DREs. These results are in line with the notion that ARNT binds DNA in the absence of AhR. Together, these results indicate that some ARNT-only peaks might also be overlapping sub-threshold peaks of AhR.

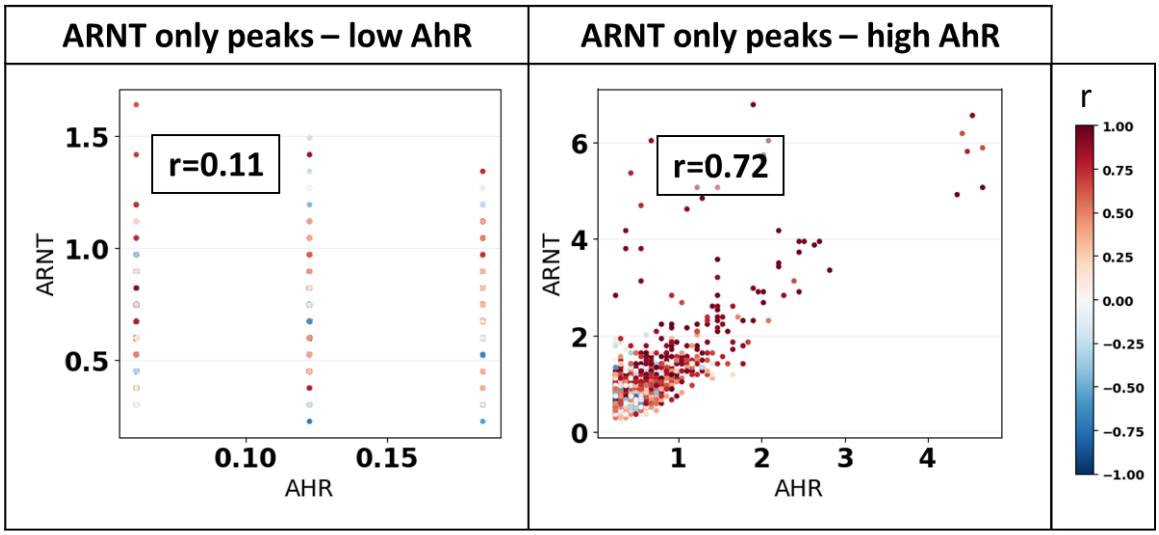

Figure 47 – AhR-ARNT interactions in TCDD-treated MCF-7 cells across ARNT-only peaks – low vs. high AhR signal.

AhR-ARNT interactions in non-treated HepG2 cells across all AhR peaks. It is currently not known whether non-activated or endogenously activated AhR interacts with TFs other than ARNT and to what extent. By focusing on investigating AhR-ARNT interactions by using the list of approximately 15,000 AhR peaks in HepG2 cells, I observed that the scatterplot correlation was very low, r=0.09 (left panel Figure 48), comparable to the results for non-interacting TFs. The proportion of peaks with signal correlations exceeding 70% (r>0.7) was 10%, and the histogram was also flat, which was also comparable to the results for non-interacting TFs (right panel Figure 48). These results suggest that ARNT might not be the primary dimerization partner of AhR in non-treated or endogenously treated cells.

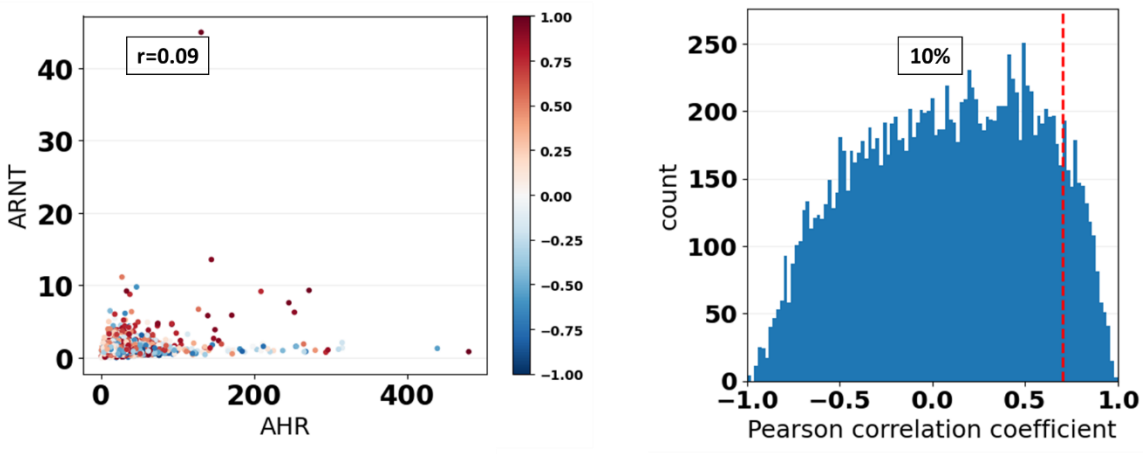

Figure 48 – AhR-ARNT interactions in non-treated HepG2 cells across all AhR peaks.

AhR-ARNT interactions in non-treated HepG2 cells across a subset of AhR peaks. I found that there were only 75 AhR peaks (about 0.5% of all AhR peaks) with more than 1 DRE and with high correlation between AhR and ARNT signals (r>0.9). The scatterplot correlation coefficient was much higher for this subset – r = 0.7 (Figure 49). One of these peaks contains two DREs and was located in the upstream region of CYP1A1, approximately -1kb from the transcription start site (TSS).

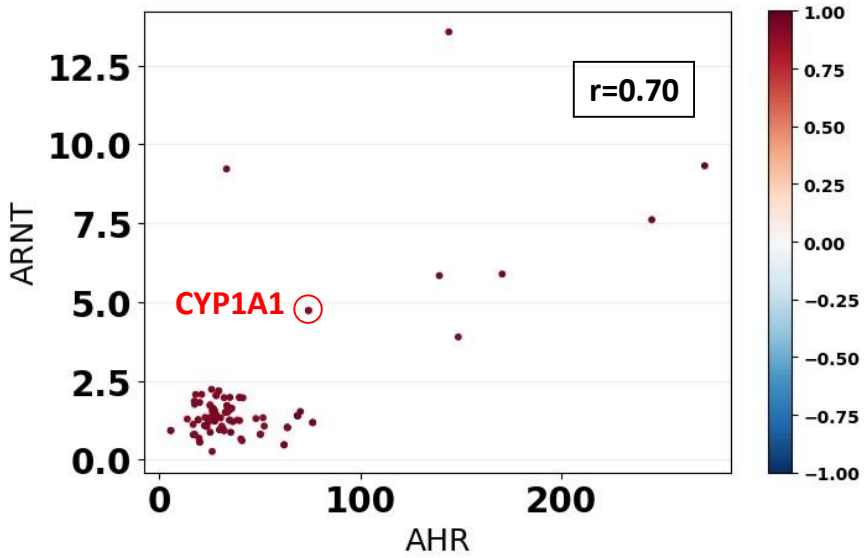

Figure 49 – AhR-ARNT interactions in non-treated HepG2 cells across a subset of AhR peaks.

AhR interactions with RelA

To confirm and further investigate AhR interactions with RelA, I examined one pair of AhR and RelA experiments (Table 4), with both AhR and RelA experiments not treated by an AhR ligand in HepG2 cell line and available on ENCODE.

Cell line	AhR	RelA
HepG2	No treatment	No treatment

Table 4 – AhR and RelA ChIP-seq experiment list.

AhR-RelA interactions in HepG2 cells across all AhR peaks. Here I compared the binding of an untreated AhR experiment and an untreated RelA experiment in HepG2 cells. The scatterplot correlation was r=0.68 (left panel Figure 50), and the proportion of AhR peaks with signal correlations exceeding 70% (r>0.7) was 34%, with the histogram slightly shifted to the left (right panel Figure 50).

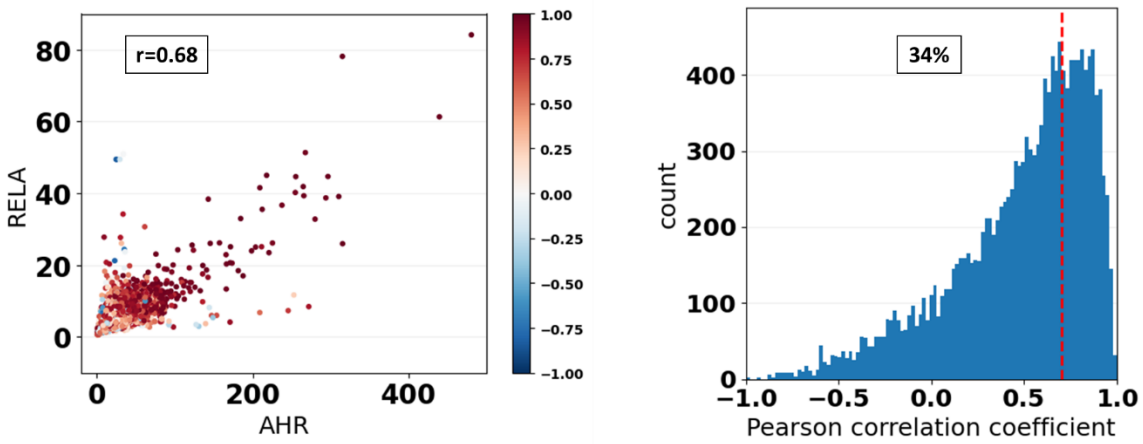

Figure 50 – AhR-RelA interactions in HepG2 cells across all AhR peaks.

AhR-RelA interactions in HepG2 cells across AhR peaks with 0 and 3+ DREs. Next, I investigated how the DRE, influenced the AhR-RelA interaction results. I have again generated the scatterplots and histograms of TF interactions, but this time for **(1)** AhR peaks with exactly 0 DREs (left panel Figure 51), and for **(2)** AhR peaks with 3 or more DREs – 3+ DREs (right panel

Figure 51). The results indicate that the scatterplot correlation decreased with increasing number of DREs, r=0.72 and r=0.51 (top left and right panels Figure 51). The proportion of peaks with signal correlations exceeding 70% (r>0.7) was 35% and 34%, for 0 DRE and 3+ DREs AhR peaks, respectively (bottom left and right panels Figure 51). Interestingly, the correlation between AhR and RelA binding decreased with increasing number of DREs under the peak, however the percentage of highly correlated peaks (peaks with r>0.7) remains the same.

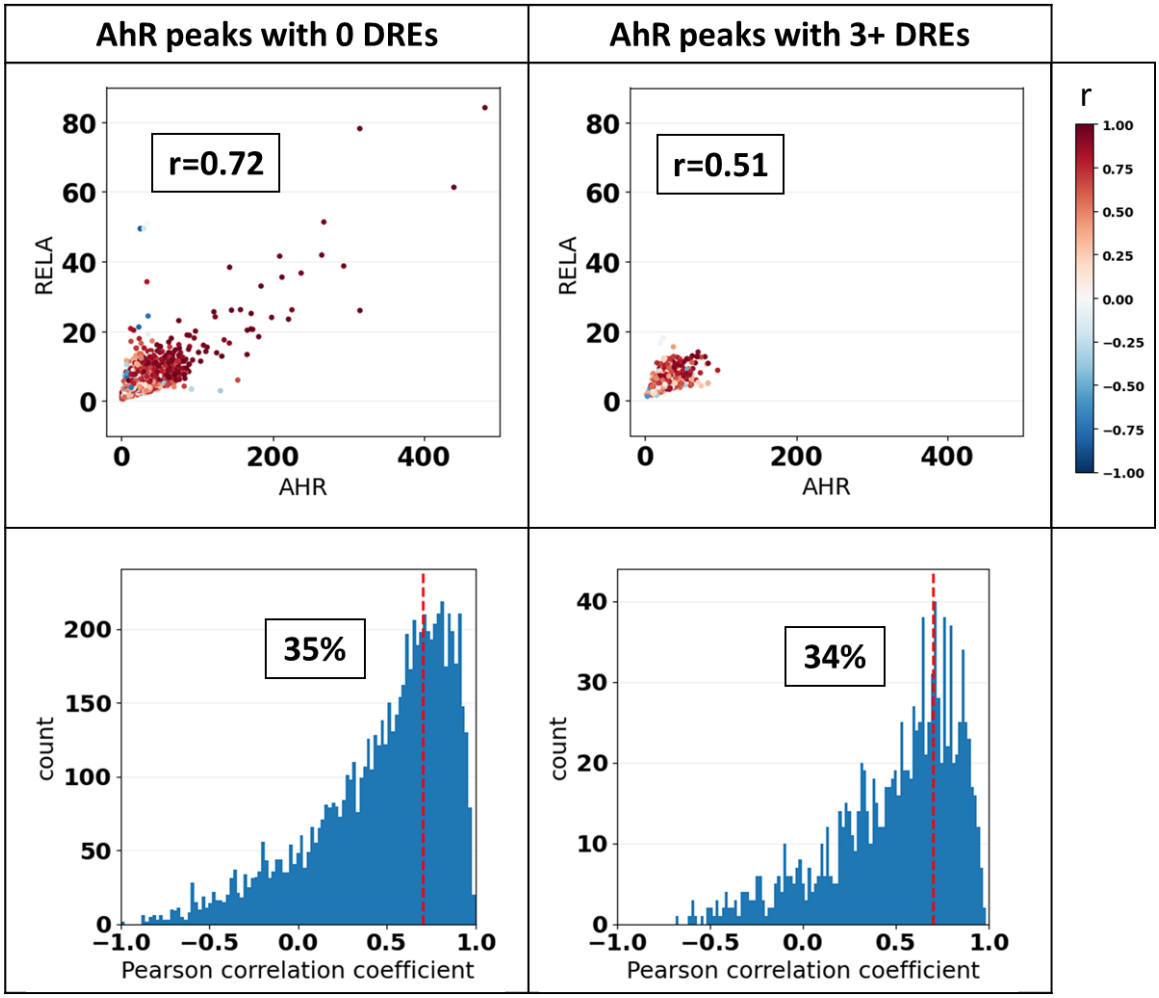

Figure 51 – AhR-RelA interactions in HepG2 cells across AhR peaks with 0 and 3+ DREs.

AhR-RelA interactions in HepG2 cells across AhR peaks with high peak to peak signal correlation between AhR and RelA. Next, I investigated the scatterplot correlation between AhR and RelA, but only across AhR peaks that have high peak to peak signal correlation

between AhR and RelA (peaks with signal to signal correlation of r>0.9), i.e., the right-most portion of the histogram of individual signal correlations. The scatterplot correlation increased to r=0.89 (Figure 52) compared to r=0.62 across all AhR peaks. Taken together, these results indicate that AhR likely interacts with RelA at a subset of AhR peaks.

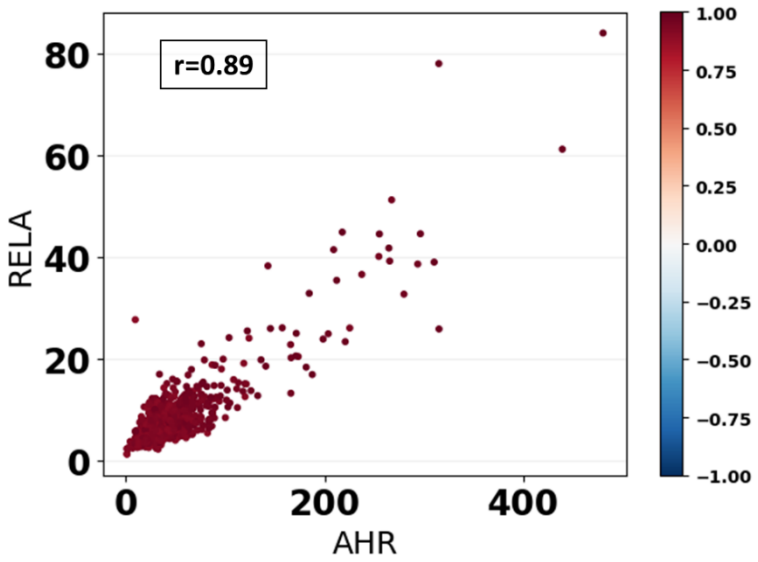

Figure 52 – AhR-RelA interactions in HepG2 cells across AhR peaks with high peak to peak signal correlation between AhR and RelA.

AhR interactions with KLF6

To confirm and further investigate AhR interactions with KLF6, I examined one pair of AhR and KLF6 experiments (Table 5). Both of these experiments were carried out in HepG2 cells that were not treated by an AhR ligand (data available on ENCODE portal).

Cell line	AhR	KLF6
HepG2	No treatment	No treatment

Table 5 – AhR and KLF6 ChIP-seq experiment list.

AhR-KLF6 interactions in HepG2 cells across all AhR peaks. Here I compared the binding in untreated HepG2 cells between an AhR and a KLF6 binding experiment. The scatterplot correlation was r=0.77 (left panel Figure 53), higher than the scatterplot correlation for KLF6. The

proportion of AhR peaks with signal correlations exceeding 70% (r>0.7) was 33%, and the histogram was slightly shifted to the left (right panel Figure 53), similar to RelA results.

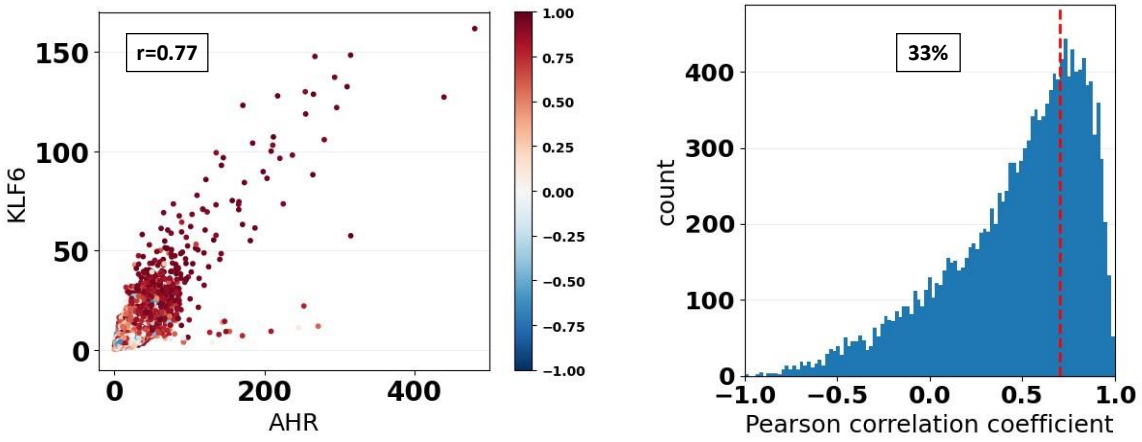

Figure 53 – AhR-KLF6 interactions in HepG2 cells across all AhR peaks.

AhR-KLF6 interactions in HepG2 cells across AhR peaks with 0 and 3+ DREs. Next, I investigated how the number of DREs under the AhR peak influenced the AhR-KLF6 interaction results. I have again generated the scatterplots and histograms of TF interactions, but this time for **(1)** AhR peaks with exactly 0 DREs (left panel Figure 54), and for **(2)** AhR peaks with 3 or more DREs – 3+ DREs (right panel Figure 54). The scatterplot correlation decreased with increasing number of DREs, r=0.81 and r=0.65 (top left and right panels Figure 51). The proportion of peaks with signal correlations exceeding 70% (r>0.7) was 33% and 34%, for 0 DRE and 3+ DREs AhR peaks, respectively (bottom left and right panels Figure 51). Similar to interactions with RelA, the correlation between AhR and RelA binding decreased with increasing number of DREs under the peak, however the percentage of highly correlated peaks (peaks with r>0.7) remained the same.

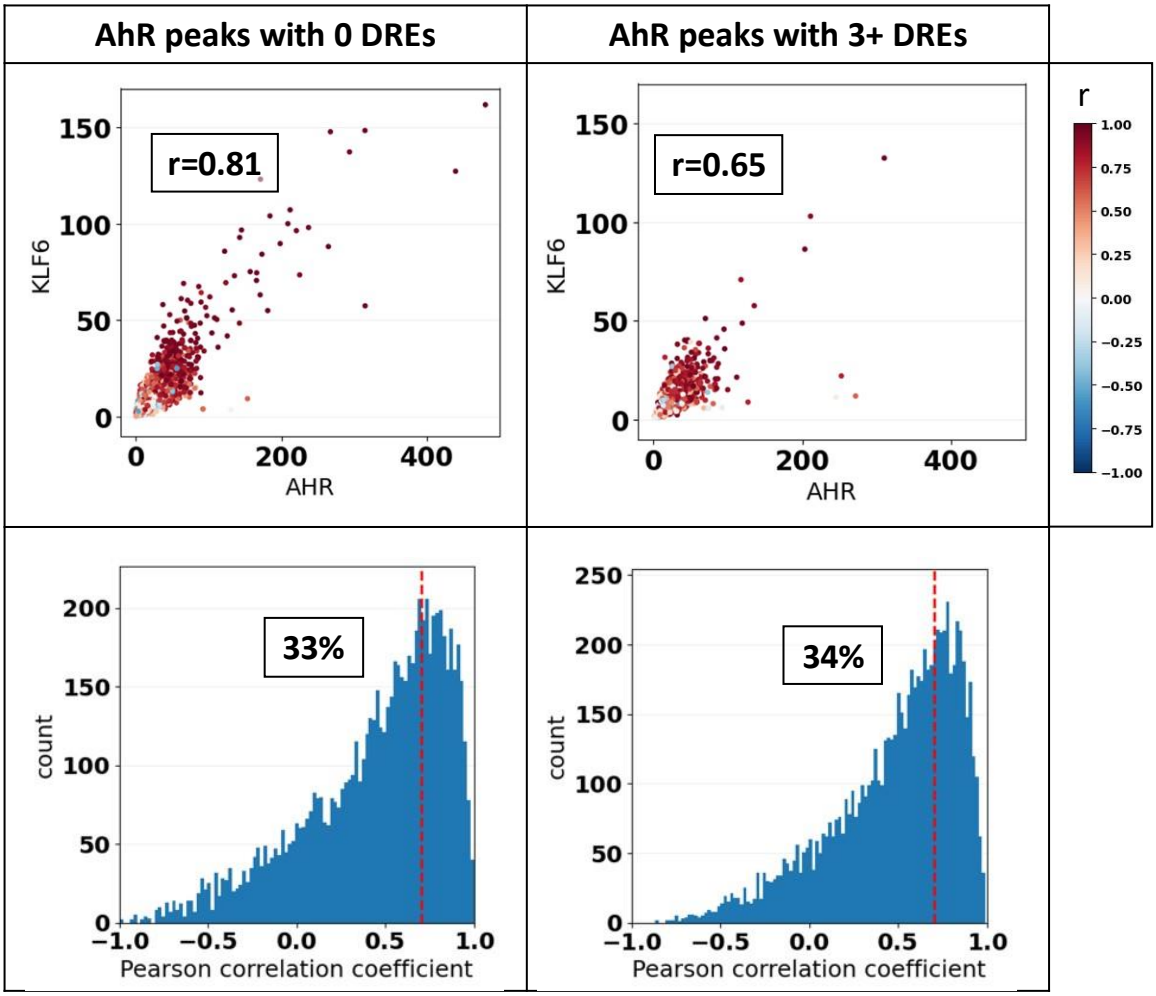

Figure 54 – AhR-KLF6 interactions in HepG2 cells across AhR peaks with 0 and 3+ DREs.

AhR-KLF6 interactions in HepG2 cells across AhR peaks with high peak to peak signal correlation between AhR and KLF6. Next, I investigated the scatterplot correlation between AhR andKLF6, but only across AhR peaks that have high peak to peak signal correlation between AhR and KLF6 (peaks with signal to signal correlation of r>0.9), i.e., the right-most portion of the histogram of individual signal correlations. The scatterplot correlation increased to r=0.85 (Figure 55) compared to r=0.77 across all AhR peaks. These results jointly indicate that AhR likely interacts with KLF6 at a subset of AhR peaks and that this subset might be slightly larger than the subset of AhR peaks where AhR and RelA interact. An implication of these results is that at some AhR peaks, AhR likely interacts with both RelA and KLF6 at the same time.

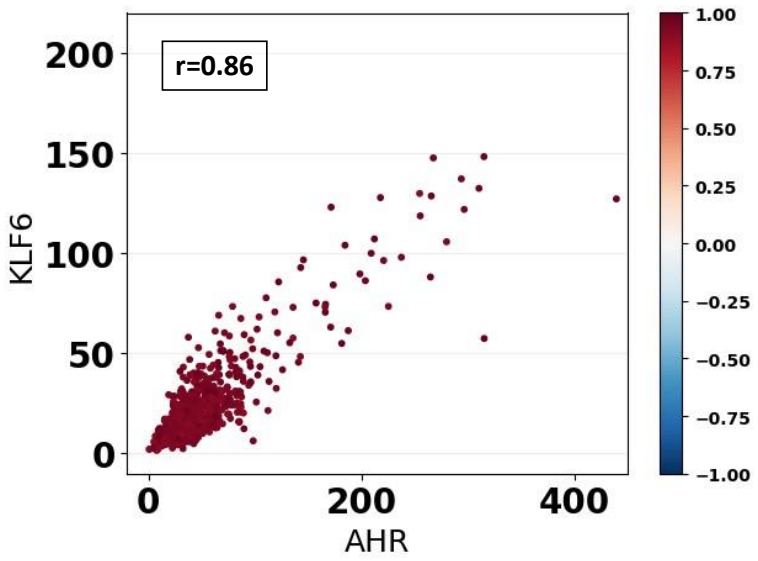

Figure 55 – AhR-KLF6 interactions in HepG2 cells across AhR peaks with high peak to peak signal correlation between AhR and KLF6.

Taking the genomic sequence of 0-DRE AhR peaks where AhR and KLF6 exhibit high peak to peak signal correlation (r>0.9) in the +/- 200-bp region around the mid-point of each peak, I ran the MEME-ChIP motif discovery pipeline. Surprisingly, the most enriched motif was the RE1 silencing transcription factor (REST) motif, shown in Figure 56. In addition to being the most enriched motif, it was also highly centrally enriched. REST is not known to interact with either AhR, KLF6 or RelA.

Figure 56 – de novo motif discovery across 0-DRE AhR peaks with high KLF6 correlation in HepG2 cells.

AhR-KLF6 interactions in HepG2 cells across AhR peaks without the REST motif. Since the role of REST in AhR-KLF6 interactions is unknown and unexpected, I investigated the correlation between AhR and KLF6 across AhR peaks that did not possess a REST motif. To achieve this, I first searched for the REST motif, JASPAR motif MA0138.2 (103), under AhR peaks, looking at sequences within the region of +/-200 from the mid-point of the peak. I used

FIMO with a q-value cutoff of 0.001 (104). This search generated 841 AhR peaks without the REST motif. The scatterplot correlation was lower than for all AhR peaks, r=0.56 (left panel Figure 57) compared to r=0.77 for all AhR peaks. The proportion of peaks with signal correlations exceeding 70% (r>0.7) was 33% (right panel Figure 57), similar to result obtained for all AhR peaks.

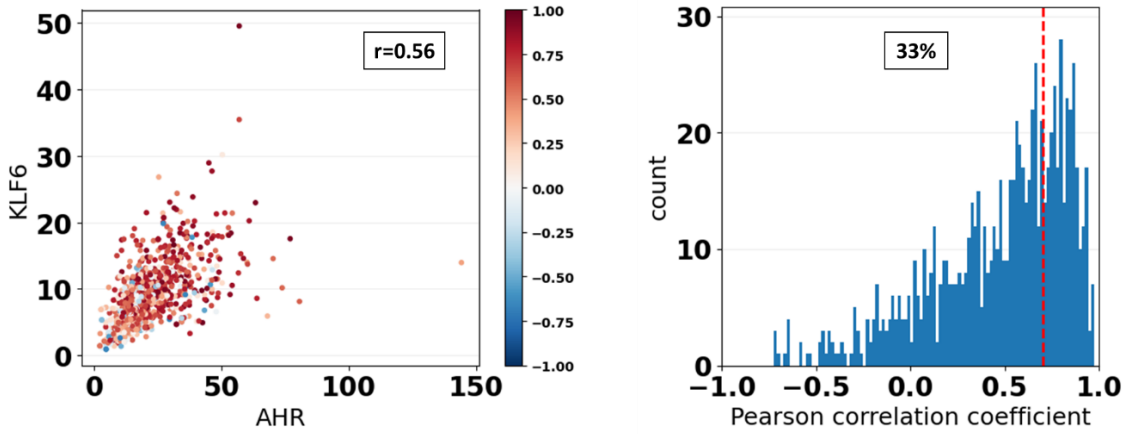

Figure 57 – AhR-KLF6 interactions in HepG2 cells across AhR peaks without the REST motif.

Performing another de novo motif search across AhR peaks without the REST motif, I found the KLF6-like motif under these peaks (Figure 58). The KLF6 motif contains little information overall and is generally degenerate, so exact matching with high confidence is difficult. The q-value reported by TomTom for the match between the KLF6 and the found motif was 0.0213. Notably, nucleotides at positions 5 and 11 do not match well between the two motifs.

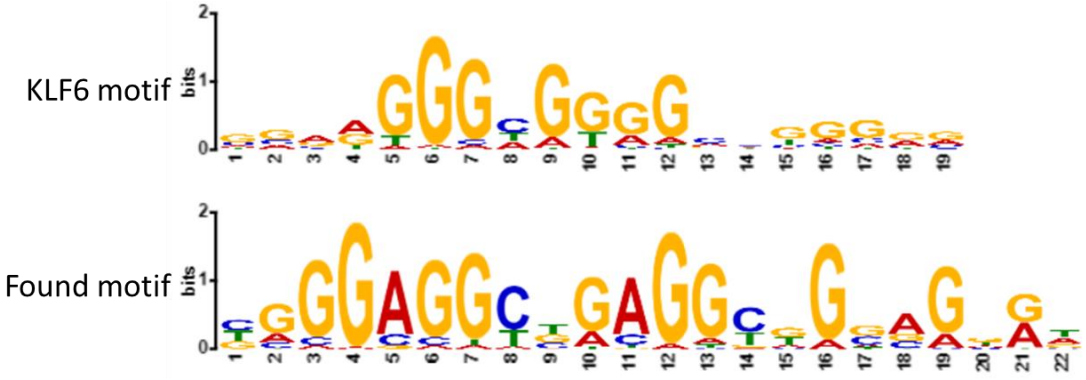

Figure 58 – de novo motif discovery across AhR peaks without the REST motif in HepG2 cells.

The de novo discovered motif looks remarkably like the sequence used to probe the NC-DRE. Huang and Elferink identified a potential binding site for AhR by examining a species-conserved sequence in the promoter of plasminogen activator inhibitor 1 (PAI-1) (99). They subsequently ran electronic mobility shift assays (EMSA) to probe the extent of AhR binding. They tested the wild-type (WT) sequence and five sequence mutants (labeled M1, M2, M3, M4 and M5). Mutants M2 through M5 lie in the portion of the sequence overlapping the de novo found motif. The motif, together with the WT and all the mutant sequences is shown in Figure 59. The mutated portions of the sequence in the mutant sequences are labelled by a red line drawn on top of the mutated sequence.

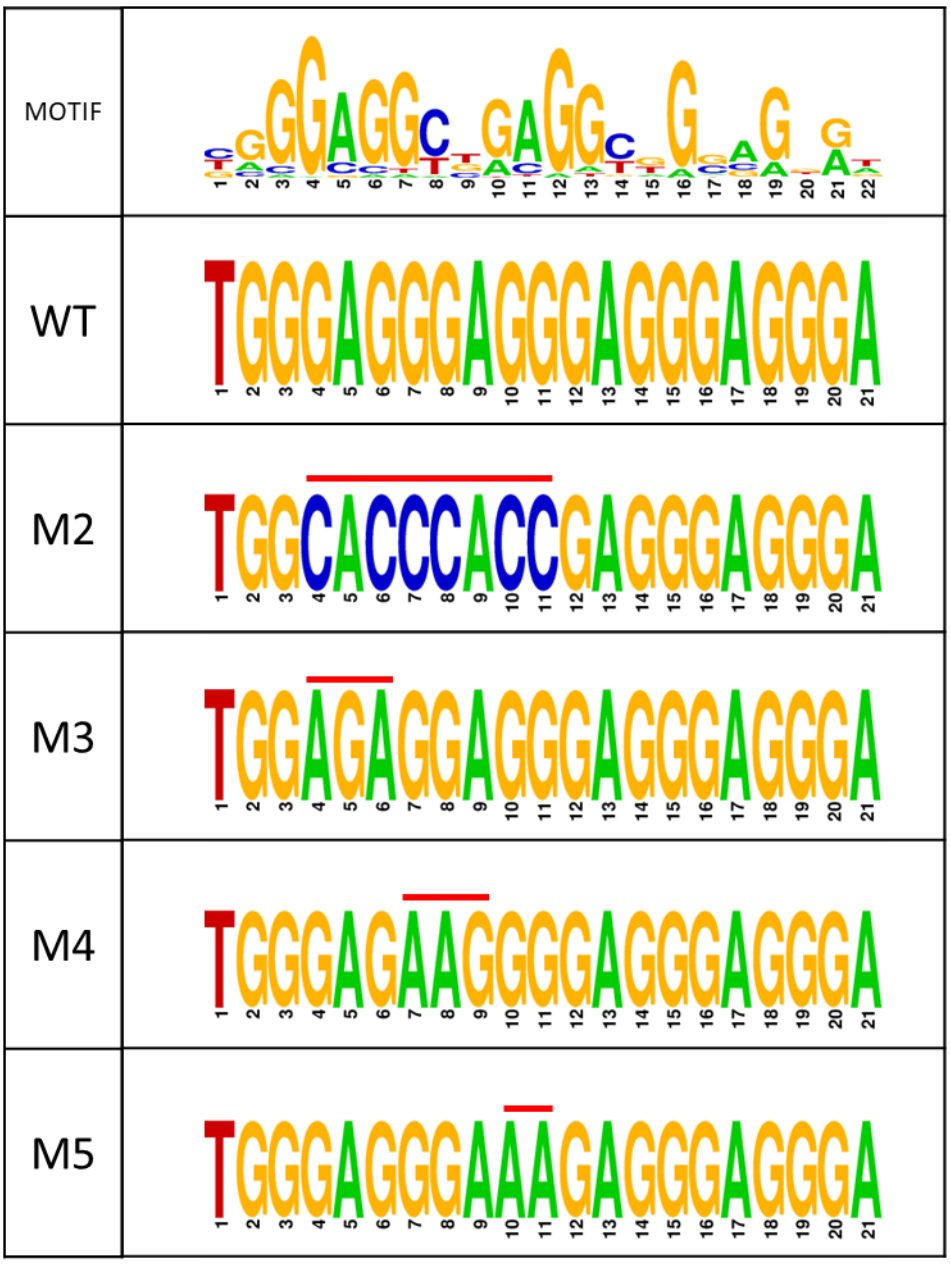

Figure 59 – comparison of de novo derived motif and functionally tested sequences in PAI-1 promoter.

Next, I calculated the position specific scoring matrix (PSSM) of the de novo discovered motif using Biopython (105), with pseudocounts calculated using the *motifs.jaspar.calculate_pseudocounts* function. I applied that PSSM to calculate the score for the WT and each of the mutant sequences (Figure 60).

Sequence	Motif score
WT	6.37
M2	-4.78
M3	-5.32
M4	7.00
M5	8.92

Figure 60 – PSSM scores of the putative AhR-KLF6 motif for five PAI-1 sequences.

These results are only in partial agreement with the functional analysis performed by Huang and Elferink (99). In their in vitro experiments mutants M2 and M4 exhibited impaired binding, while M3 and M5 did not. Further, in their in vivo experiments, mutants M3 and M5 exhibited the ability to activate a luciferase promoter, whereas mutant M4 did not (mutant M2 was not tested). Motif scores for mutants M2 and M5 are in line with previous work, while scores for mutants M3 and M4 are not. To match the functional analysis results, the score of the M3 sequence should be higher and the score of the M4 sequence should be lower. Admittedly, with only 70 sequences used to construct the motif, I suspect that there is not enough power to resolve all binding sites properly. It is possible that the true motif is less sensitive to certain nucleotide alterations and more sensitive to others, which could explain the functional results of Huang and Elferink. The consensus sequence of the motif is TGGGAGGCTGAGGCGGGAGGG, and the score for this sequence is 27.82.

CHAPTER 4: BISPHENOL A AND BISPHENOL S PREGNANCY-SPECIFIC PHYSIOLOGICALLY-BASED TOXICOKINETIC MODELS

INTRODUCTION

Bisphenols are a large class of chemicals structurally identified as having two hydroxyphenyl rings. Many bisphenols are considered endocrine disrupting chemicals (EDCs) (106). They are widely used in the manufacturing of polycarbonate plastics, epoxy resins, dental sealants, and plastic and paper consumer products (31, 107), and are pervasively present in dust and soil (32, 33). Due to consumer concerns and heightened regulations regarding the use of bisphenol A (BPA) in some countries (108), industrial and consumer products producers have resorted to using less studied bisphenol alternatives in their products (109). Such BPA-alternatives include bisphenol S (BPS), which is structurally similar to BPA, and is becoming just as environmentally prevalent (110). As a consequence, BPS is the second leading bisphenol found in humans following BPA (107, 111, 112). Bisphenols can be detected in urine, blood, breast milk, amniotic fluid and cord blood, highlighting the ubiquitous exposure humans have to these chemicals (32, 111–116). Several studies have shown that even at low concentrations, exposure to BPA during gestation can result in negative effects on the development of the fetus (117, 118). The detection of BPS in human fetal cord blood (119), the positive association between BPS exposure and prolonged gestational length (120), and the fact that in mammals, fetal exposure to BPS can alter reproductive (121, 122), metabolic (123), and behavioral outcomes (124), warrant research into the precise toxicokinetic mechanisms of these emerging bisphenol chemicals during pregnancy.

Physiologically based toxicokinetic (PBTK) mathematical models integrate toxicokinetic processes such as chemical absorption, distribution, metabolism, and excretion (ADME). The main advantage of PBTK models over the classical compartmental approaches to understanding chemical toxicokinetics is the ability of PBTK models to extrapolate outside of the conditions or population that was evaluated experimentally (125). The quantitative predictive and extrapolative capabilities of PBTK models can inform health risk assessments for chemical and pharmaceutical exposure (126, 127). Chemical toxicokinetics during pregnancy are more complex with the inclusion of the maternal, placental, and fetal compartments (126). Moreover, ethical constraints do not allow for any toxicokinetic studies other than biomonitoring to be conducted in pregnant women. The use of refined fetal surgery techniques in a sheep animal model represents unique opportunities to monitor the maternal, amniotic, and fetal compartments; key elements of pregnancy-specific PBTK (p-PBTK) models (126). Importantly, sheep are excellent models to study placental function (128, 129) and have been used for the study of feto-maternal transfer of drugs (130, 131) and EDCs (34, 132), as they allow for the simultaneous and longitudinal characterization of the pregnancy multi-compartment model in real time.

The toxicokinetics of BPA have been extensively studied and modeled in both animals and humans (133–138). Primary metabolism (conjugation) for BPA occurs in the liver and the intestine (139). In rodents, BPA undergoes substantial enterohepatic recirculation. However, in monkeys and humans, the rapid metabolism and extensive renal excretion of BPA metabolites means that a negligible amount of conjugated BPA is able to undergo enterohepatic recirculation (140, 141). In pregnancy, both conjugation and deconjugation reactions also occur in fetal tissues, primarily the fetal liver, but these processes occur at varying rates during different developmental windows. In the early developmental stages deconjugation dominates with conjugation barely occurring (142).

However, in the case of BPA, conjugation has been shown to increase from 512-fold lower to 13-fold lower when compared with maternal conjugation rates from early to late pregnancy (132).

Despite the breadth of work on BPA, only a limited number of studies have investigated the toxicokinetics of BPS during pregnancy (35, 36). Of the two available BPS toxicokinetic models (134, 143) only one is physiologically-based (134) and it is based on a non-pregnant sheep dataset. This non-pregnant BPS model was derived by a substitution of parameters from a previously calibrated BPA PBTK model with parameter values derived from quantitative structure–activity relationships (QSARs) for BPS, but was neither formally calibrated, nor validated. Recently, BPS was reported to reach higher systemic concentrations than BPA in humans (144), representing a need to better distinguish toxicokinetic characteristics between bisphenols, for which PBTK models are uniquely suited. Therefore, the objective of my current study was to improve the understanding of pregnancy toxicokinetics for bisphenols through the development of physiologically relevant multi-compartment p-PBTK models for BPA and BPS. Both p-PBTK models were developed using three independent pair-matched maternal and fetal sheep exposure cohort datasets (34–36). The text and figures in this chapter have been published as a research paper and are reprinted here with the permission of the publisher (30).

MATERIALS AND METHODS

Datasets

Experimental datasets used in this work were obtained from previously published bisphenol toxicokinetic studies in pregnant sheep (34–36). For model calibration, three independent datasets were used (two for each bisphenol). Dataset #1 from (35) , reported total (conjugated plus unconjugated) bisphenol concentrations for BPA and BPS in the maternal and fetal plasma and was used for calibrating both bisphenol models. In brief, toxicokinetic data was

obtained from pregnant Polypay × Dorset sheep (singleton pregnancies only) that underwent fetal catheterization surgery at gestational day (GD) 115. Females (n = 3) were injected with a single subcutaneous dose of BPS (0.5 mg/kg) or a combination of BPA and BPS (n = 3; 0.5 mg/kg for each chemical) and data were collected over a 72-h period. No differences in toxicokinetic parameters (maximum concentration reached, time of maximum concentration, half-life, area under the curve, area under the first moment curve, mean residence time, and total body clearance) between single-chemical exposure and mixture dosing were reported, so all BPS values (n = 6) were combined.

Additionally, two other toxicokinetic studies in pregnant sheep which presented data for conjugated and unconjugated bisphenols (34, 36) were used during model calibration. For BPA, dataset #2 was obtained from (34), who used pregnant Lacaune sheep (unreported fetal number) that underwent fetal catheterization surgery between GD 108 and 117. In separate experiments, females (n = 8) and fetuses (n = 8, unreported sex) were dosed with an intravenous (IV) infusion over 24 h of unconjugated BPA or BPA-glucuronide (conjugated, BPA-G) at a dose of 2.0 and 3.54 mg/kg/day respectively in the mother, and 5.0 and 3.54 mg/kg/day respectively in the fetus, assuming a 2.5 kg fetus. Plasma concentrations were collected over a 46-h period and the steady state plasma concentration over the final 3-h of infusion was reported. For BPS, dataset #3 was obtained from (36) which included pregnant Lacaune sheep (unreported fetal number) that underwent fetal catheterization surgery between GD 109 and 113. A dual dosing strategy was used, where pregnant females (n = 8) and their fetuses received simultaneous IV doses. First the mother received a dose of 2.7 mg/kg BPS-glucuronide (conjugated, BPS-G) and the fetus was administered a dose of 5 mg deuterated BPS (BPS-d8). This procedure was followed by a simultaneous administration of 5 mg/kg BPS to the mother and 17.5 mg BPS-G-d8 to the fetus.

Plasma concentrations were reported over a 72-h period. Dataset #3 was collected at somewhat regular intervals, though not always at the exact same time point (36). As such, these data could not be directly aggregated to yield mean and standard deviation values. Instead, the plasma data for each animal was interpolated using a cubic spline. The most representative time points were selected, and all the interpolated time-concentration curves were sampled at the selected time points and aggregated together. The time points used were either those containing the most data points across animals for all time points except the first and the last ones, or time points lying within each sheep's interpolation region, for the first and last timepoint, to prevent extrapolation.

Model development

To establish informative and useful p-PBTK models for BPA and BPS, I developed a minimal generic p-PBTK model for an unconjugated bisphenol and its conjugate metabolite that includes 6 compartments for the mother (liver, fat, kidney, placenta, blood and rest of the body) and 3 compartments for the fetus (liver, blood and the rest of the body) (Figure 61). All relevant biological processes were included, namely conjugation (metabolism) in the maternal and fetal livers, maternal urinary and biliary excretion, and deconjugation in the fetal liver (145). The two coupled sub-models of identical structure for the unconjugated and conjugated bisphenols were connected through liver metabolism in the mother and the fetus, as well as deconjugation in the fetal liver, with one sub-model used for the parent compound (BPA or BPS) and another for the conjugate (BPA_{conj} or BPS_{conj}). In the case of BPS, a duplicate model was developed for deuterated BPS and BPS_{conj} to account for fetal administration. Subsequently, I determined the physiological parameters for an average pregnant sheep, with a single fetus, at the gestational age where the experimental data were generated. This was done for the generic model, as well as for separately parametrized and calibrated individual instances of the generic model for both unconjugated BPA

and BPS, and their respective conjugated metabolites. I based the model structure and types of compartments and processes to be inclusive of the only two, to my knowledge, published BPA p-PBTK models (135, 137). All compartments were considered perfusion limited for both unconjugated and conjugated bisphenol sub-models. The most common bisphenol conjugate is glucuronide, although others, such as sulfate, exist (146). Due to a lack of available data on non-glucuronide conjugates, all conjugates for each parent compound were combined into a single conjugate parameter (BPA_{conj}/BPS_{conj}) which was calibrated against glucuronide-conjugate data.

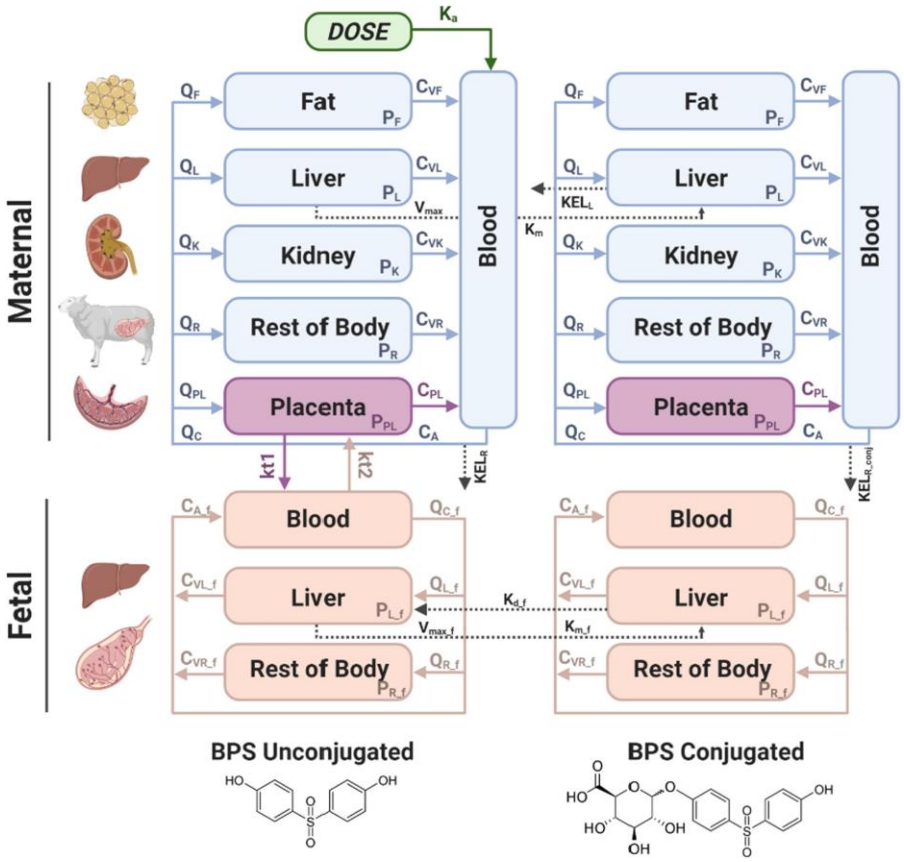

Figure 61 – PBTK model schematic.

Model equations

The equations listed in this section describe both the BPA and the BPS p-PBTK models and are the same for both compounds. The model equations for the maternal unconjugated bisphenol models are described below (Figure 62). All transport equations were perfusion limited.

Equations for conjugation represented saturable metabolism in both maternal and fetal livers and the equation for deconjugation in the fetal liver, as well as maternal urinary and biliary excretion equations were modeled as first order processes. All physiological and biochemical parameter units can be found in Table 6, Table 7, and Table 9.

$$V_K \frac{dC_K}{dt} = Q_K \cdot \left(C_A - \frac{C_K}{P_K} \right)$$

$$V_L \frac{dC_L}{dt} = Q_L \cdot \left(C_A - \frac{C_L}{P_L} \right) - \frac{V_{max} \cdot C_L/P_L}{(K_m + C_L/P_L)}$$

$$V_F \frac{dC_F}{dt} = Q_F \cdot \left(C_A - \frac{C_F}{P_F} \right)$$

$$V_R \frac{dC_R}{dt} = Q_R \cdot \left(C_A - \frac{C_R}{P_R} \right)$$

$$V_B \frac{dC_A}{dt} = \sum Q_{T\_all} \cdot \frac{C_{T\_all}}{P_{T\_all}} - Q_C \cdot C_A - KEL_R \cdot C_A + F_{SC} \cdot K_a \cdot A_{SC}$$

$$V_{PL} \frac{dC_{PL}}{dt} = Q_{PL} \cdot (C_A - C_{VPL}) - kt_1 C_{VPL} + kt_2 C_{A\_f}$$

$$C_{VT} = \frac{C_T}{P_T}$$

Figure 62 – model equations for the maternal unconjugated bisphenols.

V_T is the volume of tissue T, Q_T is the blood perfusion, C_T is the chemical concentration, P_T is the blood:tissue partition coefficient, and C_{VT} is the concentration in the venous blood exiting the tissue. C_A is the chemical concentration in the arterial blood. kt_1 and kt_2 are the diffusion rates from maternal placental blood to fetal blood and fetal blood to maternal placental blood, respectively. Subscript *_f* denotes fetal tissues. T_all is used in the maternal blood compartment to describes the sum of all tissue compartments. F_{SC} and A_{SC} represent the bioavailability of subcutaneous administration and remaining unabsorbed subcutaneous dose, respectively, and K_a is the first order rate constant for subcutaneous absorption. KEL_R is the rate of renal excretion. V_{max} is the maximum reaction rate, and K_m the Michaelis-Menten constant.

The model equations for the maternal conjugated bisphenol models are described as follows (Figure 63). Equations that are the same as in the unconjugated bisphenol models have been omitted.

$$V_L \frac{dC_L{}^{(c)}}{dt} = Q_L \left(C_A{}^{(c)} - \frac{C_L{}^{(c)}}{P_L{}^{(c)}} \right) + \frac{V_{max} \cdot \frac{C_L}{P_L}}{\left(K_m + \frac{C_L}{P_L} \right)} - KEL_L{}^{(c)} C_{VL}{}^{(c)} V_L$$

$$V_B \frac{dC_A{}^{(c)}}{dt} = \sum Q_{T\_all} \cdot \frac{C_{T\_all}{}^{(c)}}{P_{T\_all}{}^{(c)}} - Q_C \cdot C_A{}^{(c)} - KEL_R{}^{(c)} \cdot C_A{}^{(c)}$$

Figure 63 – model equations for the maternal conjugated $^{(c)}$ bisphenols.

Here, (c) in superscript denotes the conjugated compound, KEL_L is the rate of biliary excretion. All the other symbols have the same meaning as in the unconjugated bisphenol models.

The model equations for the fetal unconjugated bisphenol models are described below (Figure 64).

$$V_{Lf} \frac{dC_{Lf}}{dt} = Q_{Lf} \cdot \left(C_{Af} - \frac{C_{Lf}}{P_L} \right) - \frac{V_{max.f} \cdot C_{Lf}/P_L}{\left(K_{mf} + C_{Lf}/P_L \right)} + K_{df} C_{Lf}{}^{(c)} / P_L{}^{(c)}$$

$$V_{R\_f} \frac{dC_{R\_f}}{dt} = Q_{R\_f} \cdot \left(C_{A\_f} - \frac{C_{R\_f}}{P_{R\_f}} \right)$$

$$V_{Bf} \frac{dC_{Af}}{dt} = \sum Q_{Tf.all} \cdot \frac{C_{Tf.all}}{P_{Tf.all}} - Q_{Cf} \cdot C_{Af} + kt_1 C_{VPL} - kt_2 C_{Af}$$

Figure 64 – model equations for fetal unconjugated bisphenol.

Maternal liver partition coefficient (P_L) was used in the fetus, as well.

The equations for the conjugated bisphenol models in the fetus are described below (Figure 65).

$$V_{Lf} \frac{dC_{Lf}{}^{(c)}}{dt} = Q_{Lf} \left(C_{Af}{}^{(c)} - \frac{C_{Lf}{}^{(c)}}{P_L{}^{(c)}} \right) + \frac{V_{max.f} \cdot C_{Lf}/P_L}{(K_{mf} + C_{Lf}/P_L)} - K_{df} C_{Lf}{}^{(c)} / P_L{}^{(c)}$$

$$V_{B.f} \frac{dC_{A.f}{}^{(c)}}{dt} = \sum Q_{T.f.all} \cdot \frac{C_{T.f.all}{}^{(c)}}{P_{T.f.all}{}^{(c)}} - Q_{C.f} \cdot C_{A.f}{}^{(c)}$$

Figure 65 – model equations for fetal conjugated bisphenol.

Parameter	Abbreviation	Value	Units
Body weight[1]	BW	76.25	kg
Total cardiac output[2]	Q_{CC}	6.9	L/h/kg BW
Fractional blood flow to fat[3]	Q_{FC}	8.5	%
Fractional blood flow to kidney[3]	Q_{KC}	17	%
Fractional blood flow to liver[2]	Q_{LC}	18.3	%
Fractional blood flow to placenta[4]	Q_{PLC}	8	%
Fractional volume of fat[3]	V_{FC}	0.168	L/kg BW
Fractional volume of kidney[3]	V_{KC}	0.0046	L/kg BW
Fractional volume of liver[3]	V_{LC}	0.016	L/kg BW
Fractional volume of blood[3]	V_{BC}	0.057	L/kg BW
Fractional volume of feto-placental unit[1]	V_{PLEFC}	0.078	L/kg BW
Fractional volume of fetus[4]	V_{EFC}	0.0525	L/kg BW

Values listed obtained from references: [1](35), [2](147), [3](148), [4](149)

Table 6 – physiological parameters in pregnant sheep.

Parameter	Value	Units
Dose	$0.5*10^{-3}$	g/kg BW
K_a-BPS	0.183	L/h
MW-BPS	250.3	g/mol
MW-BPS-G	426	g/mol
K_a-BPA	0.204	L/h
MW-BPA	228.3	g/mol
MW-BPA-G	404	g/mol

Values listed obtained from references: (35), (134)
BW: body weight, G: glucuronide, Ka: absorption rate constant, MW: molecular weight.

Table 7 – BPA and BPS physicochemical parameters.

Parametrization

The generic bisphenol model was first partially parametrized with the pregnant sheep physiological parameters obtained from the literature (Table 6), inclusive of fractional blood flows and organ volumes. Following this procedure, two separate model instances were created for BPA and BPS using their respective physiochemical parameters (Table 7) and the tissue:blood partition coefficients (Table 8), which were calibrated within ranges of one order of magnitude around values either obtained from the literature (35, 134, 147–150), or estimated from the available log octanol:water partition parameters for compounds with similar partitioning (143, 151, 152) and calibration was performed within those ranges. Partition coefficients for the rest of the body for both BPA and BPS models were calibrated within the minimum and maximum values for all other tissues. All physiological parameters were assumed to be time-invariant due to the nature of the experimental data, which was collected over a short period of time during mid-late pregnancy.

Compartment	Abbreviation	BPS	BPS_{conj}	BPA	BPA_{conj}
Adipose	P_F	0.031	0.0027	1.160	0.220
Kidney	P_K	0.017	0.0049	0.858	3.180
Liver	P_L	2.300	2.4700	4.350	6.760
Rest of Body (maternal)	P_R	0.013	0.0019	0.044	0.154
Placenta	P_{PL}	0.106	0.0020	0.880	0.680
Fetal Rest of Body (fetal)	P_{R_f}	0.005	0.1680	0.006	0.500

Table 8 – passive biochemical parameters (tissue/blood partition coefficients).

Calibration

The calibration for both the BPA and the BPS models was carried out in four steps: (1) fetal conjugated bisphenol calibration, (2) maternal conjugated bisphenol calibration, (3) maternal complete calibration, and (4) feto-placental transfer and fetal complete calibration. Maternal body weight used was dependent on which of the three experimental datasets the model was being calibrated against. During the fetal conjugated bisphenol calibration, the appropriate fetal conjugated bisphenol IV administration experiment was used to partially calibrate the fetal model, namely the conjugated bisphenol partition parameters for the fetal liver and the rest of the body. The maternal conjugated bisphenol calibration relied on the maternal conjugated bisphenol IV administration data and was used to partially calibrate the maternal model, namely the remaining conjugated bisphenol partition coefficients (maternal kidney, fat, and rest of body), as well as urinary and biliary excretion rates. The complete maternal calibration relied on the maternal unconjugated bisphenol IV and total bisphenol subcutaneous administration data from all three datasets. These were used to fully calibrate the maternal model, namely the unconjugated bisphenol partition coefficients, and metabolism and urinary excretion rates (Table 9). During this step of the calibration, the feto-placental transfer of the unconjugated bisphenol was not accounted for to minimize the number of calibrated parameters. Feto-placental transfer and total fetal calibration relied on the maternal unconjugated bisphenol IV and subcutaneous administration to

fully calibrate the feto-placental diffusion rates and the remainder of the fetal parameters. These parameters were mainly unconjugated bisphenol partition coefficients for the rest of the body and metabolism and deconjugation rates in the fetal liver. Except for the rest of body partition coefficient, all other partition coefficients corresponding to the same tissue between the mother and the fetus were assumed equal.

Parameter	Abbreviation	BPS	BPS_{conj}	BPA	BPA_{conj}
Bioavailability (%)	F_{SC}	43	———	12.9	———
Biliary excretion (h^{-1})	KEL_L	———	0.061	———	2.052
Renal excretion (L/h)	KEL_R	0.023	4.093	0.035	0.375
Maternal Michaelis-Menten constant (mg/L)	K_m	4.79	———	3.46	———
Maternal maximum rate of metabolism ($mg/h/kg^{0.75}$)	V_{max}	8,185.66	———	3,458.40	———
Placental to fetal transfer	kt_1	0.075	———	6.733	———
Fetal to placental transfer	kt_2	0.113	———	5.412	———
Fetal deconjugation (L/h)	K_{d_f}	———	2.80	———	1.41
Fetal Michaelis-Menten constant (mg/L)	K_{m_f}	2.74	———	6.85	———
Fetal maximum rate of metabolism ($mg/h/kg^{0.75}$)	V_{max_f}	1,000.26	———	4,312.04	———

Table 9 – rate constants.

Our p-PBTK models required the use of blood-to-plasma partition coefficients as parameters, since the experimentally derived calibration datasets reported plasma concentrations. The blood-to-serum partition coefficient for BPA in rats has been experimentally determined as 1.10 (153), and blood-to-plasma partition coefficient for BPA and BPA-glucuronide in humans has been computationally estimated to be 1.05 and 0.83, respectively (154). Since calibrating the blood:plasma partition coefficient in the BPA model for both conjugated and unconjugated BPA within the range of 0.80 to 1.20 did not affect the model results in a significant way, blood:plasma partition coefficients for both conjugated and unconjugated BPA were fixed to 1, simplifying the modeling procedure. I observed similar results for BPS, and have thus fixed the blood:plasma partition coefficients of both conjugated and unconjugated BPS to 1.

Calibration of unknown parameter values was performed using sequential least square quadratic programming with random restart (155). Sequential least squares quadratic programming is a formal optimization technique known to perform well for systems requiring constrained nonlinear optimization, which was the case for my developed models. Here, each calibration procedure was repeated 500 times, each time starting from a randomly selected point within the allowable ranges of the calibrated parameters. The calibration most closely matching the datasets, using the lowest mean absolute percentage error score as the selection criteria, was chosen as the final calibration.

Extrapolation of maternal and fetal body burdens

Dosing regimens simulating daily repeated maternal and fetal exposures to both BPA and BPS were run with the calibrated ovine models using the reference dose for BPA set by the U.S. Environmental Protection Agency (50 μg/kg/day) (156). Simulations were run over a two-week period.

Computing software

The current model was coded in, and all simulations run using the Python programming language and the Python package Tellurium version 2.1.5 developed for reproducible dynamical modeling of biological networks (157). The full model code is available at https://github.com/BhattacharyaLab/BisphenolPBTK

Sensitivity analysis

Global sensitivity analyses of the fetal plasma compartment kinetics for both unconjugated and conjugated BPA and BPS were performed to identify the most influential parameters determining fetal bisphenol kinetics. Sensitivity analysis was performed using the variance-based Sobol method (158), as implemented within the SALib python library (159). Parameters determining fetal kinetics were examined between 50% and 150% of the nominal values listed in Table 8 and Table 9, and were sampled using the Saltelli sampling scheme with N = 1,000 generated samples (160). To examine simulated fetal kinetics with both a loading (absorption) and an elimination phase, subcutaneous dosing from Dataset #1 was selected, as described in *Datasets*. The sensitivity analysis was repeated every half-hour for 48 h of simulation time (excluding 0 h). The parameters included in the sensitivity analysis were the fetal hepatic metabolism parameters (V_{max_f} and K_{m_f}), deconjugation rate constant (K_{d_f}) and rest of body partition coefficients ($PR_{\_f}$ and $PR_{\_f}^{(c)}$). Additionally, I repeated the sensitivity analysis by also adding the feto-placental transfer parameters (kt_1 and kt_2).

RESULTS

Calibration

Simulations of the fully calibrated BPA model for both maternal and fetal compartments were compared to experimental dataset #1 following a single subcutaneous administration of BPA to the mother (Figure 66A maternal compartment, and Figure 66D fetal compartment). A full simulation was also performed for dataset #2 following either a 24-h IV infusion of BPA and BPA-G to the mother (Figure 66B and Figure 66C, respectively), or 24-h IV infusion of BPA and BPA-G to the fetus (Figure 66E and Figure 66F, respectively). All simulations matched the experimental data ± one standard deviation from the individual data points for total, unconjugated and conjugated BPA (34, 35).

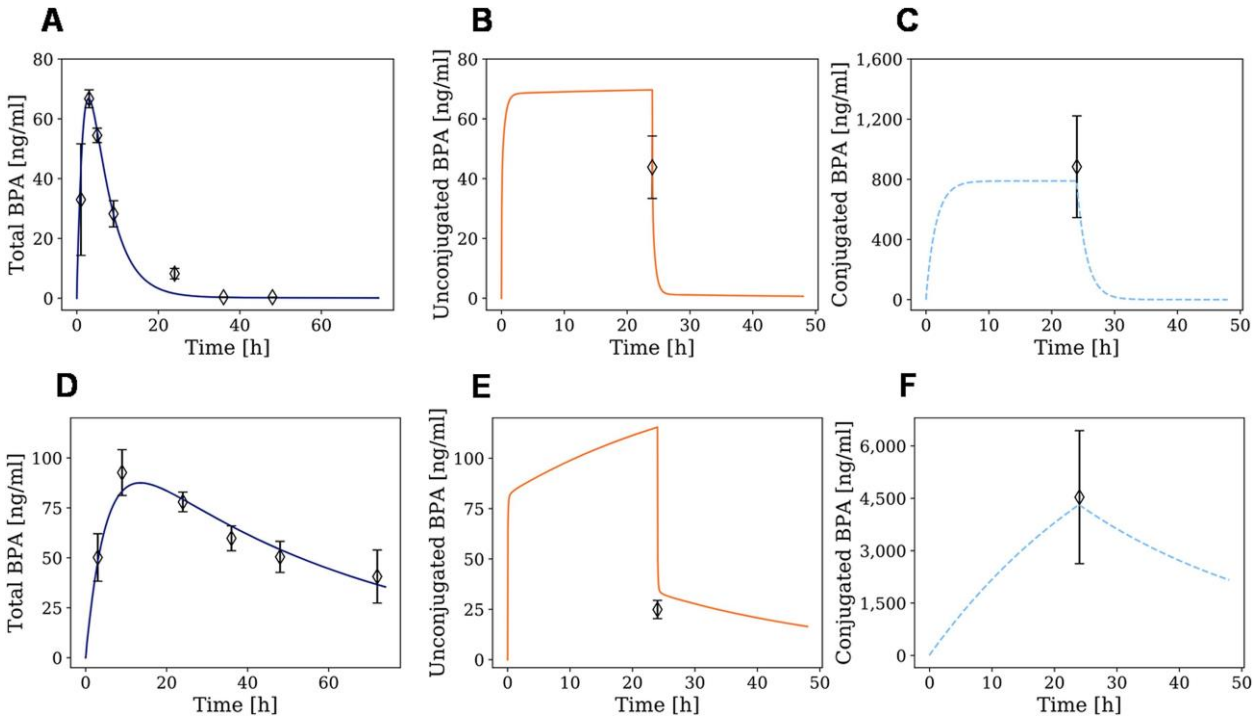

Figure 66 – simulated toxicokinetic plots of BPA for maternal and fetal circulation.

Similar to BPA, simulations of the fully calibrated BPS model were compared to experimental dataset #1 following a single subcutaneous injection of BPS to the mother (Figure 67A - maternal compartment, and Figure 67D - fetal compartment), or dataset #3 following a

single IV bolus of BPA and BPA-G to the mother (Figure 67B and Figure 67C, respectively), or a single IV bolus of BPS-d8 and BPS-G-d8 to the fetus (Figure 67E and Figure 67F, respectively). Except for fetal IV boluses of BPS-d8 and BPS-G-d8, all data points were consistent with the experimental datasets ± one standard deviation from the individual data points for total, unconjugated and conjugated BPS (35, 36).

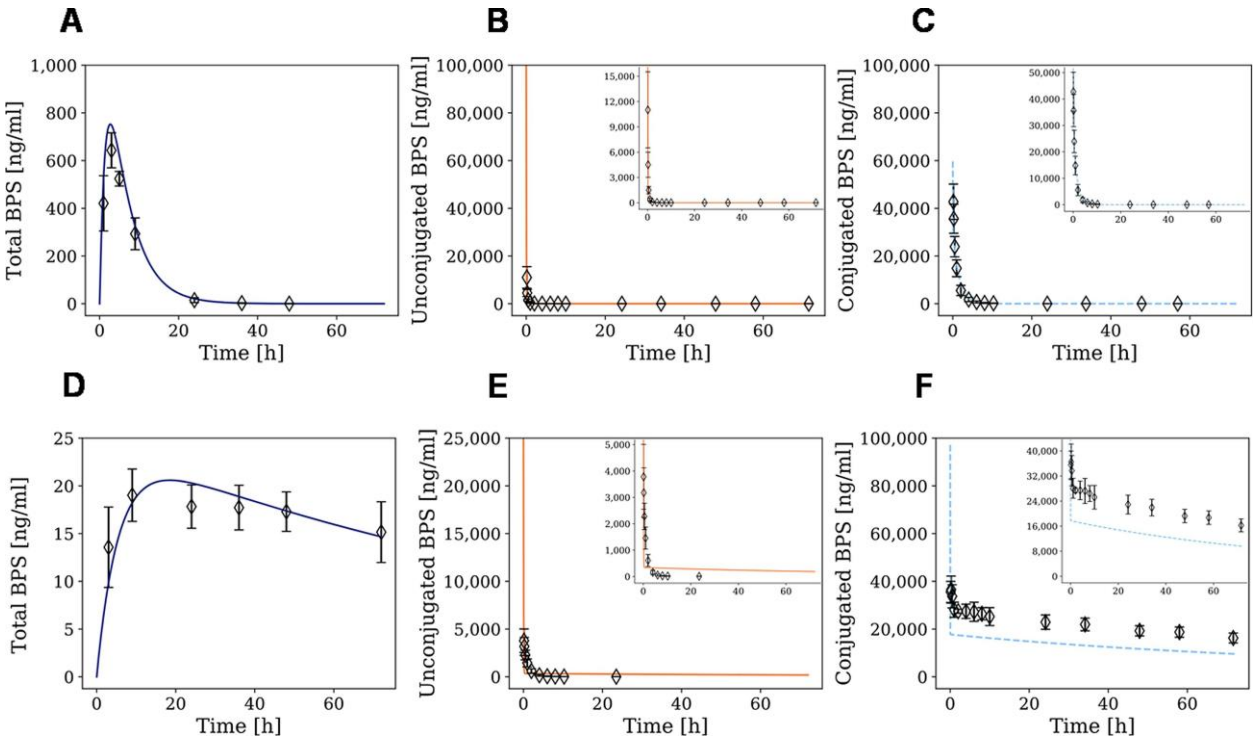

Figure 67 – simulated toxicokinetic plots of BPS for maternal and fetal circulation.

Due to its robustness, full simulations of dataset #1 (35), separated into total, conjugated, and unconjugated forms of bisphenols, were run for both BPA and BPS. This was necessary to estimate the breakdown of unconjugated and conjugated bisphenols, which was not available from the original dataset.

Extrapolation of maternal and fetal body burdens in an ovine model

Simulations showing repeated daily subcutaneous exposure to BPA and BPS are shown in Figure 68 and Figure 69, respectively. Maternal exposure was consistent with known

toxicokinetic parameters for BPA (35), where unconjugated BPA was cleared from circulation within a 24-h period (Figure 68A, *right panel*). In the fetal compartment, I observed a gradual accumulation of total, unconjugated and conjugated BPA (Figure 68B), plateauing around a mean of 0.28 ng/ml unconjugated BPA at 14 days of daily exposure (Figure 68B, *right panel, solid black line*). Like BPA, total, unconjugated and conjugated BPS also rapidly clears from maternal blood (Figure 69A) and accumulate in the fetal compartment, but total fetal BPS accumulation does not plateau within the 14-day exposure window (Figure 69B, *left panel*). The BPS model simulates fetal blood concentrations at a mean of 0.45 ng/ml unconjugated BPS by 14 days of exposure (Figure 69B, *right panel*).

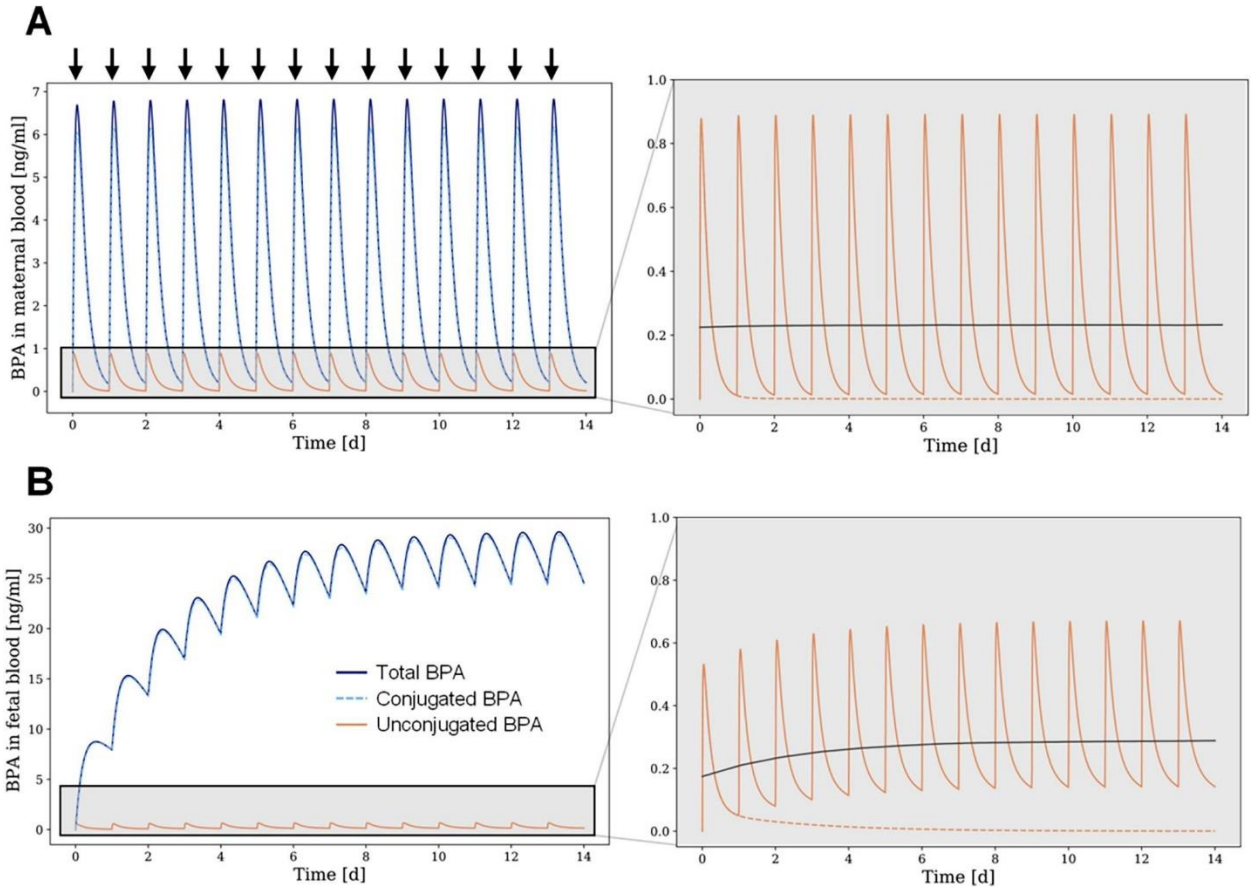

Figure 68 – simulation of BPA body burden following two weeks of daily dosing in an ovine model.

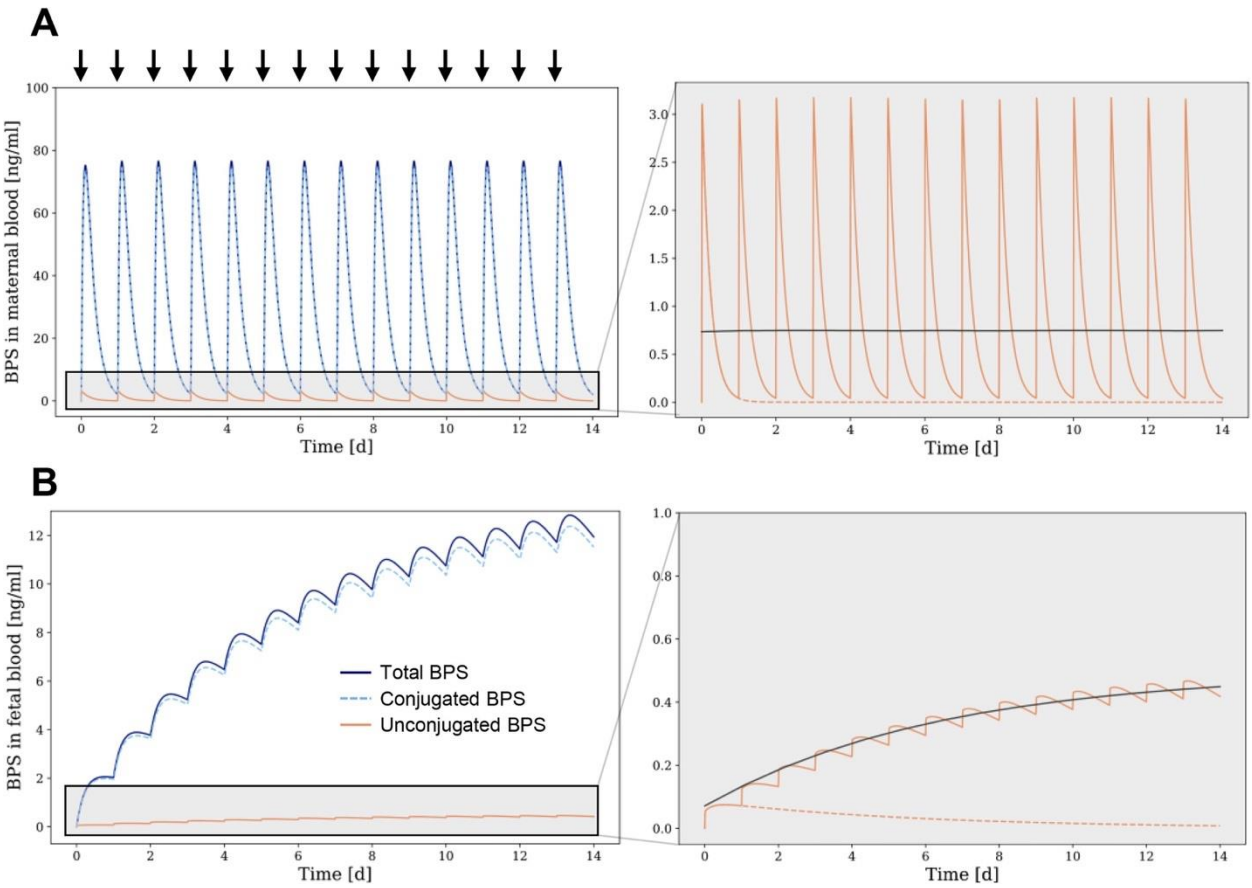

Figure 69 – simulation of BPS body burden following two weeks of daily dosing in an ovine model.

Sensitivity analysis

A global sensitivity analysis was run to investigate the main effect of all relevant fetal parameters over time and results are shown in Figure 70 (for BPA) and Figure 71 (for BPS). For both bisphenols, the main effect (%) of the placental to fetal transfer parameter kt_1 were the highest among the parameters evaluated for both unconjugated and conjugated BPA and BPS. For BPA, the main effect of the fetal to placental transfer parameter kt_2 increased over time while other parameters like fetal hepatic deconjugation (K_{d_f}) and the rate of enzymatic reaction (V_{max_f}) remained constant. For BPS, the main effect of kt_2 was lower than for BPA, but also increased over time. The contribution of other parameters that determine fetal plasma kinetics, such as metabolic (V_{max_f}, K_{m_f}) and deconjugation (K_{d_f}) parameters tended to increase over time.

Metabolic parameters (V_{max_f}, K_{m_f}) were more important for fetal plasma kinetics of unconjugated BPA, until ~ 15 h where they begin to plateau. For unconjugated BPS, the main effect of both V_{max_f} and K_{m_f} was higher than K_{d_f} throughout the 48-h period. The rest of body partition coefficient for unconjugated bisphenols ($PR_{\_f}$) had a minor contribution to output variance in determining both BPA and BPS fetal plasma kinetics, however the rest of body partition coefficient for conjugated bisphenols ($PR_{\_f}^{(c)}$) was especially important for determining conjugated BPA and BPS plasma kinetics.

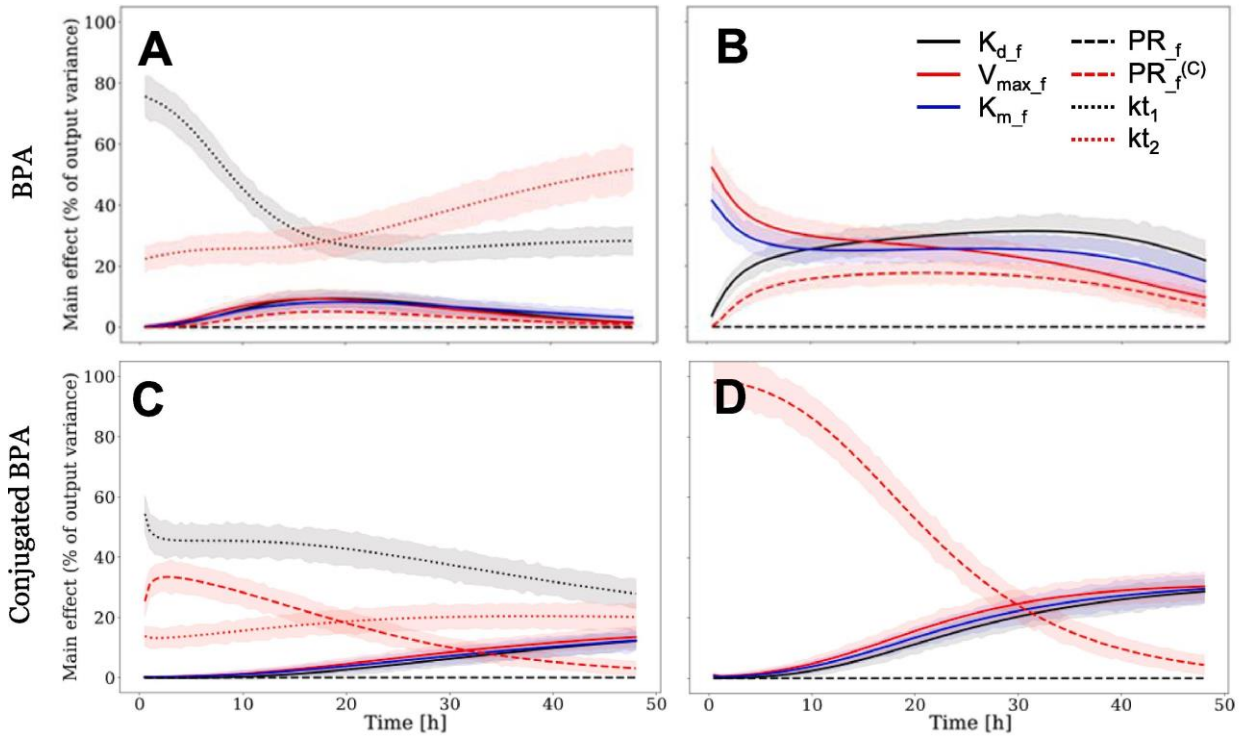

Figure 70 – global sensitivity analysis of the fetal compartment for BPA model.

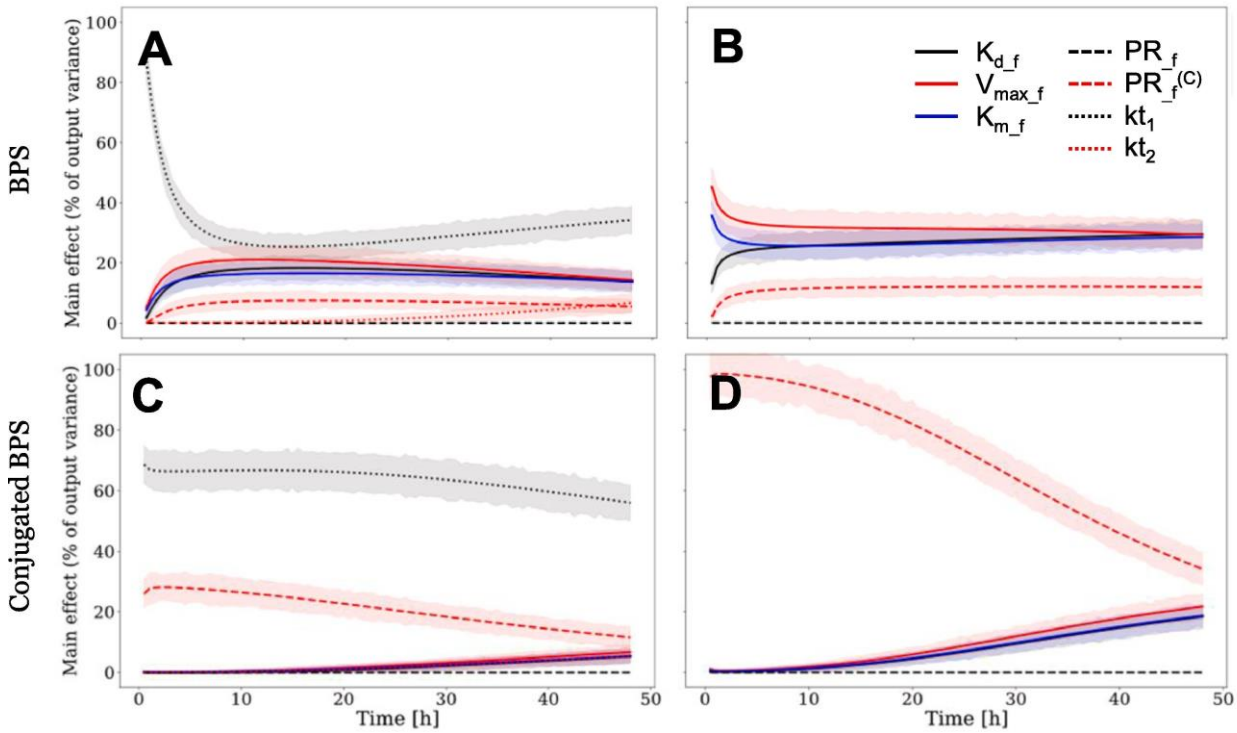

Figure 71 – global sensitivity analysis of the fetal compartment for BPS model.

CHAPTER 5: DISCUSSION AND CONCLUSIONS

Over time many computational models predicting the DNA binding of transcription factors (TFs) have been developed and the binding of specific TFs has been studied experimentally quite extensively (93, 161, 162). However, the molecular determinants and mechanisms governing the cell-specificity of binding for many TFs remain elusive. The AhR is one such TF. The AhR is a ligand-inducible TF, and its DNA binding cannot be fully determined through chromatin accessibility, the extended binding motif of AhR, the motifs of other co-bound TFs, or any combination of these features.

In vitro studies examining the DNA binding of AhR demonstrated that AhR binds exclusively to the 5'-GCGTG-3' DNA sequence, known as the dioxin response element (DRE). On the other hand, in vivo studies revealed that AhR was bound to many genomic regions that did not possess a DRE (68, 69). Thus, it is likely that some AhR-bound regions with DREs were not a result of AhR binding to those DREs. Instead, the DREs may have occurred under the AhR peaks by chance, and the AhR may have bound the DNA through some other mechanism.

My results show that many AhR peaks still have DREs and that some peaks have more DREs than could be expected by chance (Figure 3). When examining the position of the DRE within the peak relative to the mid-point of the peak, I observed that the DREs appear centrally enriched in all AhR binding experiments (Figure 4). Additionally, AhR peaks with a higher number of DREs under the peak have a statistically higher average normalized binding signal strength than AhR peaks with a lower number of DREs (Figure 9). These results suggest that DREs, although not necessary for AhR binding, are useful in determining the intensity of AhR binding.

On the other hand, AhR has also been shown capable of binding to a GC-rich region without sequence homology to the DRE. Thus, the existence of a non-canonical DRE (NC-DRE) was hypothesized. It was later demonstrated that one such NC-DRE in the promoter of plasminogen activator inhibitor 1 (PAI-1) was bound by AhR without ARNT, but together with KLF6 instead (29, 99). My analysis discovered a GC-rich motif appearing under 70 out of 841 AhR peaks having high correlation of AhR and KLF6 binding signals in the HepG2 cell line (Figure 58). This motif shares homology with the KLF6 binding motif, as well as with the identified GC-rich sequence bound by the AhR in the promoter of PAI-1. Nonetheless, the DREs still seem to play a role in HepG2 cells, as the correlation between AhR and KLF6 binding decreases with an increasing number of DREs under AhR peaks (Figure 54). Interestingly, I did not find a similar motif under AhR peaks in any other AhR binding experiment. Since the AhR binding experiment in HepG2 cells was conducted without treatment with an AhR ligand, I propose that AhR-KLF6 dimers binding to DNA do so preferentially in non-treated, or potentially endogenously treated cells.

To investigate the likely molecular determinants of AhR binding, I developed interpretable machine learning models predicting the binding status of DREs in open chromatin. These models were trained on singleton bound DREs as examples of bound DREs, but were able to predict the binding of 0-DRE AhR peaks with high accuracy. This result could be explained by a high level of indirect binding of AhR within singleton and 0-DRE AhR peaks. In this case, AhR would not be bound to the DNA directly, but instead the AhR could, for instance, be tethered to other TFs that are directly bound to DNA (161). If a sufficient number of singleton AhR peaks used in training were actually a result of such binding, then the AhR binding prediction models could learn

to recognize indirect binding of AhR and thus predict AhR binding of 0-DRE peaks. The 0-DRE AhR peaks are likely not directly bound – since they do not possess the AhR binding motif.

Another explanation for high model accuracy when predicting the binding of 0-DRE AhR peaks is that some of the bound AhR might be part of 3D chromatin loops. In this scenario, the AhR could be directly bound to one or more DREs, in one loop anchor, such that the imprint of AhR binding could also appear in other loop anchors due to their physical proximity in 3D space. EP300 transcriptional activator was one of the factors most predictive of AhR binding in enhancers (Figure 29). The EP300 has been shown to be a marker of pre-established enhancer anchors that appear in enhancer-promoter loops formed after glucocorticoid receptor (GR) activation (162). Furthermore, EP300, H3K4me1, and H3K27ac jointly mark active enhancers (163). In this scenario, the AhR molecule that was directly bound to a singleton DRE could leave an impression of a 0-DRE AhR peak in another anchor of the same loop, due to their physical proximity in 3D (48). Thus, any chromatin complexes participating in these loops would then be associated with both singleton and 0-DRE peaks. In this case my models would be learning how to identify direct AhR binding.

Interpretation of the model predictions demonstrated that binding of AhR is likely determined by 1) a common cross-cell flanking-sequence syntax (Figure 13 and Figure 38) and 2) cell-specific chromatin context syntax (Figure 12, Figure 28, Figure 29, and Figure 30). The chromatin context determinants differ vastly between cell lines or types (Figure 12, Figure 14, and Figure 36). Most commonly, one or two TFs appear to be the most important contributors to model performance in each model. I propose that these factors 1) play a significant role in determining cell identity, or 2) are a common AhR co-factor in that specific cell line or type.

The pioneering factor GATA3 is mutated in MCF-7 cells, but not in T-47D cells. The MCF-7 mutation of GATA3 is heterozygous, and results in a copy of the GATA3 protein that is more stable and resistant to turnover. Consequently, the mutated GATA3 in MCF-7 cells binds to DNA more strongly than its wildtype counterpart (164, 165). I propose that the increase in binding activity of GATA3 in MCF-7 cells makes GATA3 the most predictive factor of AhR binding in those cells. In addition, GATA3 was also shown to be the most commonly overlapping factor for binding of ERα, another inducible TF, in MCF-7 cells (166). Nonetheless, AhR peaks in T-47D cells were still correlated with GATA3 binding, albeit to a lesser extent. Therefore, wild type GATA3 might still play a role in determining AhR binding, however, this role appears to be less pronounced. Additionally, GATA3 and AhR binding have shown synergistic effects on the expression of GPR15 in human CD4+ T cells (167). Therefore, AhR-GATA3 interactions might not be confined to breast cancer cells.

On the other hand, certain TFs have been shown to be functionally associated with AhR, such as ARNT, RelA and KLF6. However, none of these factors were ranked highly by my models. ARNT is considered the principal dimerization partner of TCDD-induced AhR. However, my HepG2 and GM17212 models rank ARNT features very lowly (Figure 12). Admittedly, in the GM17212 model, the ARNT binding experiment was performed on a different but similar cell line – GM12878, and, more importantly, without AhR ligand treatment. HepG2 cells were not treated with an AhR ligand in either the AhR or the ARNT experiment. Still, when tryptophan in cell culture media is exposed to light it produces a photoproduct which has been shown to be an AhR agonist (168). In this case one could consider the ARNT experiment as having been conducted under similar conditions as the AhR experiment. Still, ARNT features were not used by the HepG2 model at all. When looking at the correlation between AhR and ARNT binding across AhR peaks

it is not difficult to see why (Figure 48). I suspect that the tryptophan derivative-induced activation of AhR might be more similar to endogenous than exogenous activation of AhR. Some of AhR's endogenous activities are likely mediated through AhR di- and multi- merization with its other known partners, such as KLF6, and RELA. Surprisingly, even though KLF6 and RELA features were used in the HepG2 model, and I have shown that the binding of AhR was highly correlated with KLF6 and RelA binding (Figure 50, Figure 53), the KLF6 and RelA features did not rank highly in feature importance. However, since I relied on a DRE-centered approach to predict AhR binding and since AhR-KLF6 dimers do not appear to directly bind DREs (Figure 58), it is possible that my selection of bound and unbound DREs in open chromatin of HepG2 cells was non-informative for the machine learning models.

In summary, I developed highly accurate and robust predictive models of within-cell line or type AhR binding. My models dissected the cell-type specificity of AhR binding and showed that cell-type specific AhR binding is driven by a complex interplay of cell-type agnostic DNA sequence immediately flanking the DRE, and a highly cell-type specific local chromatin context. Additionally, I demonstrated that ARNT was the primary binding partner of AhR in TCDD treated cells, but not in untreated cells, where KLF6 and RelA appear to be the primary binding partners of AhR.

Finally, my BPA and BPS PBTK models demonstrated chemical accumulation in the fetal compartment of a pregnant sheep experimental model; the majority of which is simulated as the bisphenol conjugate for both BPA and BPS. When considering extrapolation to daily exposure patterns in sheep, the accumulation of bisphenols in the fetal compartment has been observed in humans (169), with glucuronide conjugates being the predominant form detected (170). Using the U.S. Environmental Protection Agency's reference dose for BPA (50 μg/kg/day), I simulated

repeated maternal dosing over a two-week period for both BPA and BPS in sheep, to evaluate fetal plasma chemical burden. Here, my simulations predicted that a pseudo steady-state of 0.28 ng/ml unconjugated BPA would be reached, which falls within the range of detection for unconjugated BPA (0–53 ng/ml) in cord blood (118). For BPS, my simulations predict that a pseudo steady-state of 0.45 ng/ml unconjugated BPS would be reached. Since biomonitoring of unconjugated BPS has not been reported for cord blood, a direct comparison to human exposure cannot be made. However, the simulated total BPS in the fetal compartment (12.5 ng/ml on day 14) is in excess of total BPS concentrations measured in cord blood (<0.03–0.12 ng/ml total BPS) from a Chinese cohort (171). Most of the BPS accumulated in the fetus is predicted to be in the form of BPS conjugated metabolites. Although these metabolites are generally considered non-bioactive, BPA-G has been shown to be bioactive, and has adipogenic potential in vitro (172) . Given the predicted accumulation potential of BPS-G in the fetal compartment, the bioactivity of BPS metabolites like BPS-G should be further examined. My simulations also demonstrate that, given a steady maternal intake of BPA, unconjugated BPA rapidly reaches a state where it no longer accumulates in fetal blood. Unconjugated BPS, on the other hand, continues to accumulate in fetal blood even after 14 days of daily administrations. These results highlight the need to further study the precise fetal toxicokinetics of BPS, as well as the fetal accumulation potential of other BPA analogs.

BIBLIOGRAPHY

1. Laskowski,R.A. and Thornton,J.M. (2008) Understanding the molecular machinery of genetics through 3D structures. *Nat. Rev. Genet. 2008 92*, **9**, 141–151.

2. Latchman,D.S. (1997) Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, **29**, 1305–1312.

3. Sogawa,K. and Fujii-Kuriyama,Y. (1997) Ah Receptor, a Novel Ligand-Activated Transcription Factor. *J. Biochem.*, **122**, 1075–1079.

4. A Pharmacology Primer: Theory, Applications, and Methods - Terry Kenakin - Google Books.

5. Rothhammer,V. and Quintana,F.J. (2019) The aryl hydrocarbon receptor: an environmental sensor integrating immune responses in health and disease. *Nat. Rev. Immunol.*, **19**, 184–197.

6. Poland+,A., Glover,E. and Kende,A.S. (1976) Stereospecific, high affinity binding of 2,3,7,8-tetrachlorodibenzo-p-dioxin by hepatic cytosol. Evidence that the binding species is receptor for induction of aryl hydrocarbon hydroxylase. *J. Biol. Chem.*, **251**, 4936–4946.

7. Mimura,J. and Fujii-Kuriyama,Y. (2003) Functional role of AhR in the expression of toxic effects by TCDD. *Biochim. Biophys. Acta - Gen. Subj.*, **1619**, 263–268.

8. Nguyen,L.P. and Bradfield,C.A. (2008) The Search for Endogenous Activators of the Aryl Hydrocarbon Receptor. *Chem. Res. Toxicol.*, **21**, 102.

9. Kaiser,H., Parker,E. and Hamrick,M.W. (2020) Kynurenine signaling through the aryl hydrocarbon receptor: Implications for aging and healthspan. *Exp. Gerontol.*, **130**.

10. Perepechaeva,M.L. and Grishanova,A.Y. (2020) The Role of Aryl Hydrocarbon Receptor (AhR) in Brain Tumors. *Int. J. Mol. Sci.*, **21**.

11. Wu,D., Potluri,N., Kim,Y. and Rastinejad,F. (2013) Structure and Dimerization Properties of the Aryl Hydrocarbon Receptor PAS-A Domain. *Mol. Cell. Biol.*, **33**, 4346.

12. Pandini,A., Denison,M.S., Song,Y., Soshilov,A.A. and Bonati,L. (2007) Structural and functional characterization of the aryl hydrocarbon receptor ligand binding domain by homology modeling and mutational analysis. *Biochemistry*, **46**, 696–708.

13. Schulte,K.W., Green,E., Wilz,A., Platten,M. and Daumke,O. (2017) Structural Basis for Aryl Hydrocarbon Receptor-Mediated Gene Activation Article Structural Basis for Aryl Hydrocarbon Receptor-Mediated Gene Activation. *Struct. Des.*, **25**, 1025–1033.e3.

14. Watson,J.D., Prokopec,S.D., Smith,A.B., Okey,A.B., Pohjanvirta,R. and Boutros,P.C. (2014) TCDD dysregulation of 13 AHR-target genes in rat liver. *Toxicol. Appl. Pharmacol.*, **274**, 445–454.

15. Korashy,H.M. and El-Kadi,A.O.S. (2006) The role of aryl hydrocarbon receptor and the reactive oxygen species in the modulation of glutathione transferase by heavy metals in murine hepatoma cell lines. *Chem. Biol. Interact.*, **162**, 237–248.

16. Shimizu,Y., Nakatsuru,Y., Ichinose,M., Takahashi,Y., Kume,H., Mimura,J., Fujii-Kuriyama,Y. and Ishikawa,T. (2000) Benzo[a]pyrene carcinogenicity is lost in mice lacking the aryl hydrocarbon receptor. *Proc. Natl. Acad. Sci. U. S. A.*, **97**, 779–782.

17. Xie,G., Peng,Z. and Raufman,J.P. (2012) Src-mediated aryl hydrocarbon and epidermal growth factor receptor cross talk stimulates colon cancer cell proliferation. *Am. J. Physiol. - Gastrointest. Liver Physiol.*, **302**, G1006.

18. Rothhammer,V. and Quintana,F.J. (2019) The aryl hydrocarbon receptor: an environmental sensor integrating immune responses in health and disease. *Nat. Rev. Immunol. 2019 193*, **19**, 184–197.

19. Park,R., Madhavaram,S. and Ji,J.D. (2020) The Role of Aryl-Hydrocarbon Receptor (AhR) in Osteoclast Differentiation and Function. *Cells*, **9**.

20. Gialitakis,M., Tolaini,M., Li,Y., Pardo,M., Yu,L., Toribio,A., Choudhary,J.S., Niakan,K., Papayannopoulos,V. and Stockinger,B. (2017) Activation of the Aryl Hydrocarbon Receptor Interferes with Early Embryonic Development. *Stem Cell Reports*, **9**, 1377.

21. Wang,Z., Snyder,M., Kenison,J.E., Yang,K., Lara,B., Lydell,E., Bennani,K., Novikov,O., Federico,A., Monti,S., *et al.* (2021) How the AHR Became Important in Cancer: The Role of Chronically Active AHR in Cancer Aggression. *Int. J. Mol. Sci.*, **22**, 1–22.

22. Shimba,S. and Watabe,Y. (2009) Crosstalk between the AHR signaling pathway and circadian rhythm. *Biochem. Pharmacol.*, **77**, 560–565.

23. Marlowe,J.L. and Puga,A. (2005) Aryl hydrocarbon receptor, cell cycle regulation, toxicity, and tumorigenesis. *J. Cell. Biochem.*, **96**, 1174–1184.

24. Esser,C., Rannug,A. and Stockinger,B. (2009) The aryl hydrocarbon receptor in immunity. *Trends Immunol.*, **30**, 447–454.

25. Kobayashi,A., Sogawa,K. and Fujii-Kuriyama,Y. (1996) Cooperative interaction between AhR.Arnt and Sp1 for the drug-inducible expression of CYP1A1 gene. *J. Biol. Chem.*, **271**, 12310–12316.

26. Ge,N.L. and Elferink,C.J. (1998) A direct interaction between the aryl hydrocarbon receptor and retinoblastoma protein. Linking dioxin signaling to the cell cycle. *J. Biol. Chem.*, **273**, 22708–22713.

27. Kim,D.W., Gazourian,L., Quadri,S.A., Raphaëlle, Sherr,D.H. and Sonenshein,G.E. (2000) The RelA NF-kappaB subunit and the aryl hydrocarbon receptor (AhR) cooperate to transactivate the c-myc promoter in mammary cells. *Oncogene*, **19**, 5498–5506.

28. Chen,P.H., Chang,H., Chang,J.T. and Lin,P. (2012) Aryl hydrocarbon receptor in association with RelA modulates IL-6 expression in non-smoking lung cancer. *Oncogene*, **31**, 2555–2565.

29. Wilson,S.R., Joshi,A.D. and Elferink,C.J. (2013) The tumor suppressor Kruppel-like factor 6 is a novel aryl hydrocarbon receptor DNA binding partner. *J. Pharmacol. Exp. Ther.*, **345**, 419–429.

30. Gingrich,J., Filipovic,D., Conolly,R., Bhattacharya,S. and Veiga-Lopez,A. (2021) Pregnancy-specific physiologically-based toxicokinetic models for bisphenol A and bisphenol S. *Environ. Int.*, **147**.

31. Liao,C. and Kannan,K. (2013) Concentrations and profiles of bisphenol a and other bisphenol analogues in foodstuffs from the united states and their implications for human exposure. *J. Agric. Food Chem.*, **61**, 4655–4662.

32. Liao,C., Liu,F., Guo,Y., Moon,H.B., Nakata,H., Wu,Q. and Kannan,K. (2012) Occurrence of eight bisphenol analogues in indoor dust from the United States and several Asian countries: implications for human exposure. *Environ. Sci. Technol.*, **46**, 9138–9145.

33. Kwak,J. Il, Moon,J., Kim,D., Cui,R. and An,Y.J. (2018) Determination of the soil hazardous concentrations of bisphenol A using the species sensitivity approach. *J. Hazard. Mater.*, **344**, 390–397.

34. Corbel,T., Gayrard,V., Viguié,C., Puel,S., Lacroix,M.Z., Toutain,P.L. and Picard-Hagen,N. (2013) Bisphenol A disposition in the sheep maternal-placental-fetal unit: Mechanisms determining fetal internal exposure. *Biol. Reprod.*, **89**, 11–12.

35. Gingrich,J., Pu,Y., Ehrhardt,R., Karthikraj,R., Kannan,K. and Veiga-Lopez,A. (2019) Toxicokinetics of bisphenol A, bisphenol S, and bisphenol F in a pregnancy sheep model. *Chemosphere*, **220**, 185–194.

36. Grandin,F.C., Lacroix,M.Z., Gayrard,V., Gauderat,G., Mila,H., Toutain,P.L. and Picard-Hagen,N. (2018) Bisphenol S instead of Bisphenol A: Toxicokinetic investigations in the ovine materno-feto-placental unit. *Environ. Int.*, **120**, 584–592.

37. Sonawane,A.R., Platig,J., Fagny,M., Chen,C.-Y., Paulson,J.N., Lopes-Ramos,C.M., DeMeo,D.L., Quackenbush,J., Glass,K. and Kuijjer,M.L. (2017) Understanding Tissue-Specific Gene Regulation. *Cell Rep.*, **21**, 1077–1088.

38. Todeschini,A.L., Georges,A. and Veitia,R.A. (2014) Transcription factors: specific DNA binding and specific gene regulation. *Trends Genet.*, **30**, 211–219.

39. Caetano,M.S., Hassane,M., Van,H.T., Bugarin,E., Cumpian,A.M., McDowell,C.L., Cavazos,C.G., Zhang,H., Deng,S., Diao,L., *et al.* (2018) Sex specific function of epithelial STAT3 signaling in pathogenesis of K-ras mutant lung cancer. *Nat. Commun. 2018 91*, **9**, 1–11.

40. Warrick,J.I., Walter,V., Yamashita,H., Chung,E., Shuman,L., Amponsa,V.O., Zheng,Z., Chan,W., Whitcomb,T.L., Yue,F., *et al.* (2016) FOXA1, GATA3 and PPARɣ Cooperate to Drive Luminal Subtype in Bladder Cancer: A Molecular Analysis of Established Human Cell Lines. *Sci. Reports 2016 61*, **6**, 1–15.

41. Kress,S., Reichert,J. and Schwarz,M. (1998) Functional analysis of the human cytochrome P4501A1 (CYP1A1) gene enhancer. *Eur. J. Biochem.*, **258**, 803–812.

42. Ye,W., Chen,R., Chen,X., Huang,B., Lin,R., Xie,X., Chen,J., Jiang,J., Deng,Y. and Wen,J. (2019) AhR regulates the expression of human cytochrome P450 1A1 (CYP1A1) by recruiting Sp1. *FEBS J.*, **286**, 4215–4231.

43. Denison,M.S., Fisher,J.M. and Whitlock,J.P. (1988) The DNA recognition site for the dioxin-Ah receptor complex. Nucleotide sequence and functional analysis. *J. Biol. Chem.*, **263**, 17221–17224.

44. Swanson,H.I., Chan,W.K. and Bradfield,C.A. (1995) DNA binding specificities and pairing rules of the Ah receptor, ARNT, and SIM proteins. *J. Biol. Chem.*, **270**, 26292–26302.

45. Sun,Y. V., Boverhof,D.R., Burgoon,L.D., Fielden,M.R. and Zacharewski,T.R. (2004) Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences. *Nucleic Acids Res.*, **32**, 4512–4523.

46. Yang,S.Y., Ahmed,S., Satheesh,S. V. and Matthews,J. (2018) Genome-wide mapping and analysis of aryl hydrocarbon receptor (AHR)- and aryl hydrocarbon receptor repressor (AHRR)-binding sites in human breast cancer cells. *Arch. Toxicol.*, **92**, 225–240.

47. Lo,R. and Matthews,J. (2012) High-resolution genome-wide Mapping of AHR and ARNT binding sites by ChIP-Seq. *Toxicol. Sci.*, **130**, 349–361.

48. Liang,J., Lacroix,L., Gamot,A., Cuddapah,S., Queille,S., Lhoumaud,P., Lepetit,P., Martin,P.G.P., Vogelmann,J., Court,F., *et al.* (2014) Chromatin immunoprecipitation indirect peaks highlight long-range interactions of insulator proteins and Pol II pausing. *Mol. Cell*, **53**, 672–681.

49. Denison,M.S. and Nagy,S.R. (2003) Activation of the Aryl Hydrocarbon Receptor by Structurally Diverse Exogenous and Endogenous Chemicals. *Annu. Rev. Pharmacol. Toxicol.*, **43**, 309–334.

50. Abel,J. and Haarmann-Stemmann,T. (2010) An introduction to the molecular basics of aryl hydrocarbon receptor biology. **391**, 1235–1248.

51. Gutiérrez-Vázquez,C. and Quintana,F.J. (2018) Regulation of the Immune Response by the Aryl Hydrocarbon Receptor. *Immunity*, **48**, 19–33.

52. Perdew,G.H. (1988) Association of the Ah receptor with the 90-kDa heat shock protein. *J. Biol. Chem.*, **263**, 13802–13805.

53. Denis,M., Cuthill,S., Wikström,A.C., Poellinger,L. and Gustafsson,J.Å. (1988) Association of the dioxin receptor with the Mr 90,000 heat shock protein: A structural kinship with the glucocorticoid receptor. *Biochem. Biophys. Res. Commun.*, **155**, 801–807.

54. Carver,L.A. and Bradfield,C.A. (1997) Ligand-dependent Interaction of the Aryl Hydrocarbon Receptor with a Novel Immunophilin Homolog In Vivo. *J. Biol. Chem.*, **272**, 11452–11456.

55. Meyer,B.K. and Perdew,G.H. (1999) Characterization of the AhR-hsp90-XAP2 core complex and the role of the immunophilin-related protein XAP2 in AhR stabilization. *Biochemistry*, **38**, 8907–8917.

56. Grenert,J.P., Sullivan,W.P., Fadden,P., Haystead,T.A.J., Clark,J., Mimnaugh,E., Krutzsch,H., Ochel,H.J., Schulte,T.W., Sausville,E., *et al.* (1997) The Amino-terminal Domain of Heat Shock Protein 90 (hsp90) That Binds Geldanamycin Is an ATP/ADP Switch Domain That Regulates hsp90 Conformation. *J. Biol. Chem.*, **272**, 23843–23850.

57. Ikuta,T., Eguchi,H., Tachibana,T., Yoneda,Y. and Kawajiri,K. (1998) Nuclear Localization and Export Signals of the Human Aryl Hydrocarbon Receptor. *J. Biol. Chem.*, **273**, 2895–2904.

58. Ikuta,T., Kobayashi,Y. and Kawajiri,K. (2004) Phosphorylation of nuclear localization signal inhibits the ligand-dependent nuclear import of aryl hydrocarbon receptor. *Biochem. Biophys. Res. Commun.*, **317**, 545–550.

59. Durrin,L.K., Jones,P.B.C., Fisher,J.M., Galeazzi,D.R. and Whitlock,J.P. (1987) 2,3,7,8-Tetrachlorodibenzo-p-dioxin receptors regulate transcription of the cytochrome P1-450 gene. *J. Cell. Biochem.*, **35**, 153–160.

60. Dere,E., Lo,R., Celius,T., Matthews,J. and Zacharewski,T.R. (2011) Integration of Genome-Wide Computation DRE Search, AhR ChIP-chip and Gene Expression Analyses of TCDD-Elicited Responses in the Mouse Liver. *BMC Genomics 2011 121*, **12**, 1–19.

61. Nebert,D.W., Roe,A.L., Dieter,M.Z., Solis,W.A., Yang,Y. and Dalton,T.P. (2000) Role of the aromatic hydrocarbon receptor and (Ah) gene battery in the oxidative stress response, cell cycle control, and apoptosis. In *Biochemical Pharmacology*. Biochem Pharmacol, Vol. 59, pp. 65–85.

62. Sorg,O. (2014) AhR signalling and dioxin toxicity. *Toxicol. Lett.*, **230**, 225–233.

63. Beischlag,T. V., Morales,J.L., Hollingshead,B.D. and Perdew,G.H. (2008) The Aryl Hydrocarbon Receptor Complex and the Control of Gene Expression. *Crit. Rev. Eukaryot. Gene Expr.*, **18**, 207.

64. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, **129**, 823–837.

65. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive Genome-wide Protein-DNA Interactions Detected at Single Nucleotide Resolution. *Cell*, **147**, 1408.

66. He,Q., Johnston,J. and Zeitlinger,J. (2015) ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotechnol. 2015 334*, **33**, 395–401.

67. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet. 2004 54*, **5**, 276–287.

68. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G., *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260.

69. Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505.

70. Li,F. and Stormo,G.D. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.

71. Ogawa,N. and Biggin,M.D. (2012) High-Throughput SELEX Determination of DNA Sequences Bound by Transcription Factors In Vitro. *Methods Mol. Biol.*, **786**, 51–63.

72. Karimzadeh,M. and Hoffman,M.M. (2022) Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome. *Genome Biol. 2022 231*, **23**, 1–23.

73. Gotea,V., Visel,A., Westlund,J.M., Nobrega,M.A., Pennacchio,L.A. and Ovcharenko,I. (2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.*, **20**, 565–577.

74. Yan,J., Enge,M., Whitington,T., Dave,K., Liu,J., Sur,I., Schmierer,B., Jolma,A., Kivioja,T., Taipale,M., *et al.* (2013) Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell*, **154**, 801–813.

75. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.

76. Quang,D. and Xie,X. (2019) FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, **166**, 40–47.

77. Keilwagen,J., Posch,S. and Grau,J. (2019) Accurate prediction of cell type-specific transcription factor binding. *Genome Biol. 2019 201*, **20**, 1–17.

78. Srivastava,D. and Mahony,S. (2020) Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochim. Biophys. Acta - Gene Regul. Mech.*, **1863**, 194443.

79. Chen,T. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, pp. 785–794.

80. Filipovic,D., Qi,W., Kana,O.Z., Marri,D., LeCluyse,E.L., Andersen,M.E., Cuddapah,S. and Bhattacharya,S. (2022) Predictive Models of Genome-wide Aryl Hydrocarbon Receptor DNA Binding Reveal Tissue Specific Binding Determinants. *bioRxiv*, 10.1101/2022.05.13.491754.

81. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K., *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

82. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

83. Neavin,D.R., Lee,J.-H., Liu,D., Ye,Z., Li,H., Wang,L., Ordog,T. and Weinshilboum,R.M. (2019) Single Nucleotide Polymorphisms at a Distance from Aryl Hydrocarbon Receptor (AHR) Binding Sites Influence AHR Ligand–Dependent Gene Expression. *Drug Metab. Dispos.*, **47**, 983–994.

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

85. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.*, **19**, e46255.

86. Ramírez,F., Dündar,F., Diehl,S., Grüning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.

87. Pansoy,A., Ahmed,S., Valen,E., Sandelin,A. and Matthews,J. (2010) 3-Methylcholanthrene Induces Differential Recruitment of Aryl Hydrocarbon Receptor to Human Promoters. *Toxicol. Sci.*, **117**, 90–100.

88. S,A., E,V., A,S. and J,M. (2009) Dioxin increases the interaction between aryl hydrocarbon receptor and estrogen receptor alpha at human promoters. *Toxicol. Sci.*, **111**, 254–266.

89. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M., *et al.* (2021) The UCSC genome browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.

90. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Ridwan Amode,M., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J., *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.

91. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

92. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J., *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods 2020 173*, **17**, 261–272.

93. Elith,J., Leathwick,J.R. and Hastie,T. (2008) A working guide to boosted regression trees. *J. Anim. Ecol.*, **77**, 802–813.

94. Gregorutti,B., Michel,B. and Saint-Pierre,P. (2017) Correlation and variable importance in random forests. *Stat Comput*, **27**, 659–678.

95. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

96. Haidar,R., Henkler,F., Kugler,J., Rosin,A., Genkinger,D., Laux,P. and Luch,A. (2021) The role of DNA-binding and ARNT dimerization on the nucleo-cytoplasmic translocation of the aryl hydrocarbon receptor. *Sci. Reports 2021 111*, **11**, 1–11.

97. Sogawa,K., Nakano,R., Kobayashi,A., Kikuchi,Y., Ohe,N., Matsushita,N. and Fujii-Kuriyama,Y. (1995) Possible function of Ah receptor nuclear translocator (Arnt) homodimer in transcriptional regulation. *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 1936–1940.

98. Gassmann,M., Chilov,D. and Wenger,R.H. (2000) Regulation of the hypoxia-inducible factor-1 alpha. ARNT is not necessary for hypoxic induction of HIF-1 alpha in the nucleus. *Adv. Exp. Med. Biol.*, **475**, 87–99.

99. Huang,G. and Elferink,C.J. (2012) A Novel Nonconsensus Xenobiotic Response Element Capable of Mediating Aryl Hydrocarbon Receptor-Dependent Gene Expression. *Mol. Pharmacol.*, **81**, 338.

100. Tian,Y., Ke,S., Denison,M.S., Rabson,A.B. and Gallo,M.A. (1999) Ah receptor and NF-kappaB interactions, a potential mechanism for dioxin toxicity. *J. Biol. Chem.*, **274**, 510–515.

101. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, 1–9.

102. Kulakovskiy,I. V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A., *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252.

103. Castro-Mondragon,J.A., Riudavets-Puig,R., Rauluseviciute,I., Berhanu Lemma,R., Turchi,L., Blanc-Mathieu,R., Lucas,J., Boddie,P., Khan,A., Perez,N.M., *et al.* (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **50**, D165–D173.

104. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

105. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B., *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

106. Gore,A.C., Crews,D., Doan,L.L., Merrill,M. La, Patisaul,H. and Zota,A. (2014) Introduction to Endocrine Disrupting Chemicals (EDCs): A Guide for Public Interest Organizations and Policy Makers.

107. Liao,C., Liu,F., Alomirah,H., Loi,V.D., Mohd,M.A., Moon,H.B., Nakata,H. and Kannan,K. (2012) Bisphenol S in urine from the United States and seven Asian countries: Occurrence and human exposures. *Environ. Sci. Technol.*, **46**, 6860–6866.

108. Jalal,N., Surendranath,A.R., Pathak,J.L., Yu,S. and Chung,C.Y. (2018) Bisphenol A (BPA) the mighty and the mutagenic. *Toxicol. Reports*, **5**, 76–84.

109. EPA (2015) Bisphenol A Alternatives in Thermal Paper.

110. Rochester,J.R. and Bolden,A.L. (2015) Bisphenol S and F: A systematic review and comparison of the hormonal activity of bisphenol a substitutes. *Environ. Health Perspect.*, **123**, 643–650.

111. Philips,E.M., Jaddoe,V.W.V., Asimakopoulos,A.G., Kannan,K., Steegers,E.A.P., Santos,S. and Trasande,L. (2018) Bisphenol and phthalate concentrations and its determinants among pregnant women in a population-based cohort in the Netherlands, 2004–5. *Environ. Res.*, **161**, 562–572.

112. Ye,X., Wong,L.Y., Kramer,J., Zhou,X., Jia,T. and Calafat,A.M. (2015) Urinary Concentrations of Bisphenol A and Three Other Bisphenols in Convenience Samples of U.S. Adults during 2000-2014. *Environ. Sci. Technol.*, **49**, 11834–11839.

113. Asimakopoulos,A.G., Xue,J., De Carvalho,B.P., Iyer,A., Abualnaja,K.O., Yaghmoor,S.S., Kumosani,T.A. and Kannan,K. (2016) Urinary biomarkers of exposure to 57 xenobiotics and its association with oxidative stress in a population in Jeddah, Saudi Arabia. *Environ. Res.*, **150**, 573–581.

114. Lehmler,H.J., Liu,B., Gadogbe,M. and Bao,W. (2018) Exposure to Bisphenol A, Bisphenol F, and Bisphenol S in U.S. Adults and Children: The National Health and Nutrition Examination Survey 2013-2014. *ACS Omega*, **3**, 6523–6532.

115. Rocha,B.A., Asimakopoulos,A.G., Honda,M., da Costa,N.L., Barbosa,R.M., Barbosa,F. and Kannan,K. (2018) Advanced data mining approaches in the assessment of urinary concentrations of bisphenols, chlorophenols, parabens and benzophenones in Brazilian children and their association to DNA damage. *Environ. Int.*, **116**, 269–277.

116. Xue,J., Wu,Q., Sakthivel,S., Pavithran,P. V., Vasukutty,J.R. and Kannan,K. (2015) Urinary levels of endocrine-disrupting chemicals, including bisphenols, bisphenol A diglycidyl ethers, benzophenones, parabens, and triclosan in obese and non-obese Indian children. *Environ. Res.*, **137**, 120–128.

117. Gingrich,J., Ticiani,E. and Veiga-Lopez,A. (2020) Placenta Disrupted: Endocrine Disrupting Chemicals and Pregnancy. *Trends Endocrinol. Metab.*, **31**, 508–524.

118. Veiga-Lopez,A., Pu,Y., Gingrich,J. and Padmanabhan,V. (2018) Obesogenic Endocrine Disrupting Chemicals: Identifying Knowledge Gaps. *Trends Endocrinol. Metab.*, **29**, 607–625.

119. Kolatorova,L., Vitku,J., Hampl,R., Adamcova,K., Skodova,T., Simkova,M., Parizek,A., Starka,L. and Duskova,M. (2018) Exposure to bisphenols and parabens during pregnancy and relations to steroid changes. *Environ. Res.*, **163**, 115–122.

120. Wan,Y., Huo,W., Xu,S., Zheng,T., Zhang,B., Li,Y., Zhou,A., Zhang,Y., Hu,J., Zhu,Y., *et al.* (2018) Relationship between maternal exposure to bisphenol S and pregnancy duration. *Environ. Pollut.*, **238**, 717–724.

121. Gingrich,J., Pu,Y., Roberts,J., Karthikraj,R., Kannan,K., Ehrhardt,R. and Veiga-Lopez,A. (2018) Gestational bisphenol S impairs placental endocrine function and the fusogenic trophoblast signaling pathway. *Arch. Toxicol.*, **92**, 1861–1876.

122. Kolla,S.D.D., Morcos,M., Martin,B. and Vandenberg,L.N. (2018) Low dose bisphenol S or ethinyl estradiol exposures during the perinatal period alter female mouse mammary gland development. *Reprod. Toxicol.*, **78**, 50–59.

123. Pu,Y., Gingrich,J.D., Steibel,J.P. and Veiga-Lopez,A. (2017) Sex-Specific Modulation of Fetal Adipogenesis by Gestational Bisphenol A and Bisphenol S Exposure. *Endocrinology*, **158**, 3844–3858.

124. Catanese,M.C. and Vandenberg,L.N. (2017) Bisphenol S (BPS) Alters Maternal Behavior and Brain in Mice Exposed During Pregnancy/Lactation and Their Daughters. *Endocrinology*, **158**, 516–530.

125. Tsamandouras,N., Rostami-Hodjegan,A. and Aarons,L. (2015) Combining the 'bottom up' and 'top down' approaches in pharmacokinetic modelling: fitting PBPK models to observed clinical data. *Br. J. Clin. Pharmacol.*, **79**, 48–55.

126. Ke,A.B., Greupink,R. and Abduljalil,K. (2018) Drug Dosing in Pregnant Women: Challenges and Opportunities in Using Physiologically Based Pharmacokinetic Modeling and Simulations. *CPT Pharmacometrics Syst. Pharmacol.*, **7**, 103–110.

127. Zhuang,X. and Lu,C. (2016) PBPK modeling and simulation in drug research and development. *Acta Pharm. Sin. B*, **6**, 430–440.

128. Fowden,A.L., Forhead,A.J., Sferruzzi-Perri,A.N., Burton,G.J. and Vaughan,O.R. (2015)

Review: Endocrine regulation of placental phenotype. *Placenta*, **36**, S50–S59.

129. Mourier,E., Tarrade,A., Duan,J., Richard,C., Bertholdt,C., Beaumont,M., Morel,O. and Chavatte-Palmer,P. (2016) Non-invasive evaluation of placental blood flow: lessons from animal models. *Reproduction*, **153**, R85–R96.

130. Krishna,R., Riggs,K.W., Kwan,E., Wong,H., Szeitz,A., Walker,M.P.R. and Rurak,D.W. (2002) Clearance and disposition of indometacin in chronically instrumented fetal lambs following a 3-day continuous intravenous infusion. *J. Pharm. Pharmacol.*, **54**, 801–808.

131. Ngamprasertwong,P., Dong,M., Niu,J., Venkatasubramanian,R., Vinks,A.A. and Sadhasivam,S. (2016) Propofol Pharmacokinetics and Estimation of Fetal Propofol Exposure during Mid-Gestational Fetal Surgery: A Maternal-Fetal Sheep Model. *PLoS One*, **11**, e0146563.

132. Corbel,T., Perdu,E., Gayrard,V., Puel,S., Lacroix,M.Z., Viguié,C., Toutain,P.L., Zalko,D. and Picard-Hagen,N. (2015) Conjugation and Deconjugation Reactions within the Fetoplacental Compartment in a Sheep Model: A Key Factor Determining Bisphenol A Fetal Exposure. *Drug Metab. Dispos.*, **43**, 467–476.

133. Fisher,J.W., Twaddle,N.C., Vanlandingham,M. and Doerge,D.R. (2011) Pharmacokinetic modeling: Prediction and evaluation of route dependent dosimetry of bisphenol A in monkeys with extrapolation to humans. *Toxicol. Appl. Pharmacol.*, **257**, 122–136.

134. Karrer,C., Roiss,T., von Goetz,N., Skledar,D.G., Mašič,L.P. and Hungerbühler,K. (2018) Physiologically based pharmacokinetic (PBPK) modeling of the bisphenols BPA, BPS, BPF, and BPAF with new experimental metabolic parameters: Comparing the pharmacokinetic behavior of BPA with its substitutes. *Environ. Health Perspect.*, **126**.

135. Kawamoto,Y., Matsuyama,W., Wada,M., Hishikawa,J., Chan,M.P.L., Nakayama,A. and Morisawa,S. (2007) Development of a physiologically based pharmacokinetic model for bisphenol A in pregnant mice. *Toxicol. Appl. Pharmacol.*, **224**, 182–191.

136. Poet,T. and Hays,S. (2017) Extrapolation of plasma clearance to understand species differences in toxicokinetics of bisphenol A. *https://doi.org/10.1080/00498254.2017.1379626*, **48**, 891–897.

137. Sharma,R.P., Schuhmacher,M. and Kumar,V. (2018) The development of a pregnancy PBPK Model for Bisphenol A and its evaluation with the available biomonitoring data. *Sci. Total Environ.*, **624**, 55–68.

138. Vom Saal,F.S., Vandevoort,C.A., Taylor,J.A., Welshons,W. V., Toutain,P.L. and Hunt,P.A. (2014) Bisphenol A (BPA) pharmacokinetics with daily oral bolus or continuous exposure via silastic capsules in pregnant rhesus monkeys: Relevance for human exposures. *Reprod. Toxicol.*, **45**, 105–116.

139. Domoradzki,J.Y., Pottenger,L.H., Thornton,C.M., Hansen,S.C., Card,T.L., Markham,D.A., Dryzga,M.D., Shiotsuka,R.N. and Waechter,J.M. (2003) Metabolism and Pharmacokinetics

of Bisphenol A (BPA) and the Embryo-Fetal Distribution of BPA and BPA-Monoglucuronide in CD Sprague-Dawley Rats at Three Gestational Stages. *Toxicol. Sci.*, **76**, 21–34.

140. Doerge,D.R., Twaddle,N.C., Woodling,K.A. and Fisher,J.W. (2010) Pharmacokinetics of bisphenol A in neonatal and adult rhesus monkeys. *Toxicol. Appl. Pharmacol.*, **248**, 1–11.

141. Völkel,W., Colnot,T., Csanády,G.A., Filser,J.G. and Dekant,W. (2002) Metabolism and Kinetics of Bisphenol A in Humans at Low Doses Following Oral Administration. *Chem. Res. Toxicol.*, **15**, 1281–1287.

142. Lucier,G.W., Sonawane,B.R. and McDaniel,O.S. (1977) Glucuronidation and deglucuronidation reactions in hepatic and extrahepatic tissues during perinatal development. *Drug Metab. Dispos.*, **5**.

143. Oh,J., Choi,J.W., Ahn,Y.A. and Kim,S. (2018) Pharmacokinetics of bisphenol S in humans after single oral administration. *Environ. Int.*, **112**, 127–133.

144. Khmiri,I., Côté,J., Mantha,M., Khemiri,R., Lacroix,M., Gely,C., Toutain,P.L., Picard-Hagen,N., Gayrard,V. and Bouchard,M. (2020) Toxicokinetics of bisphenol-S and its glucuronide in plasma and urine following oral and dermal exposure in volunteers for the interpretation of biomonitoring data. *Environ. Int.*, **138**, 105644.

145. Nishikawa,M., Iwano,H., Yanagisawa,R., Koike,N., Inoue,H. and Yokota,H. (2010) Placental transfer of conjugated bisphenol A and subsequent reactivation in the rat fetus. *Environ. Health Perspect.*, **118**, 1196–1203.

146. Ho,K.L., Yuen,K.K., Yau,M.S., Murphy,M.B., Wan,Y., Fong,B.M.W., Tam,S., Giesy,J.P., Leung,K.S.Y. and Lam,M.H.W. (2017) Glucuronide and Sulfate Conjugates of Bisphenol A: Chemical Synthesis and Correlation Between Their Urinary Levels and Plasma Bisphenol A Content in Voluntary Human Donors. *Arch. Environ. Contam. Toxicol.*, **73**, 410–420.

147. Craigmill,A.L. (2003) A physiologically based pharmacokinetic model for oxytetracycline residues in sheep. *J. Vet. Pharmacol. Ther.*, **26**, 55–63.

148. Upton,R.N. (2008) Organ weights and blood flows of sheep and pig for physiological pharmacokinetic modelling. *J. Pharmacol. Toxicol. Methods*, **58**, 198–205.

149. Makowski,E.L., Meschia,G., Droegemueller,W. and Battaglia,F.C. (1968) Measurement of umbilical arterial blood flow to the sheep placenta and fetus in utero. Distribution to cotyledons and the intercotyledonary chorion. *Circ. Res.*, **23**, 623–631.

150. NCBI 4,4'-Sulfonyldiphenol, CID=6626.

151. Chow,E.C.Y., Talattof,A., Tsakalozou,E., Fan,J., Zhao,L. and Zhang,X. (2016) Using Physiologically Based Pharmacokinetic (PBPK) Modeling to Evaluate the Impact of Pharmaceutical Excipients on Oral Drug Absorption: Sensitivity Analyses. *AAPS J.*, **18**,

1500–1511.

152. Lyons,M.A., Reisfeld,B., Yang,R.S.H. and Lenaerts,A.J. (2013) A physiologically based pharmacokinetic model of rifampin in mice. *Antimicrob. Agents Chemother.*, **57**, 1763–1771.

153. Shin,B.S., Kim,C.H., Jun,Y.S., Kim,D.H., Lee,B.M., Yoon,C.H., Park,E.H., Lee,K.C., Han,S.-Y., Park,K.L., *et al.* (2004) PHYSIOLOGICALLY BASED PHARMACOKINETICS OF BISPHENOL A. *J. Toxicol. Environ. Heal. Part A*, **67**, 1971–1985.

154. Edginton,A.N. and Ritter,L. (2009) Predicting plasma concentrations of bisphenol A in children younger than 2 years of age after typical feeding schedules, using a physiologically based toxicokinetic model. *Environ. Health Perspect.*, **117**, 645–652.

155. Bonnans,J.F., Gilbert,J.C., Lemaréchal,C. and Sagastizábal,C.A. (2003) Numerical Optimization. 10.1007/978-3-662-05078-1.

156. EPA (2012) Bisphenol A, CASRN 80-05-7. IRIS (Integrated Risk Information System). Washingron, DC.

157. Medley,J.K., Choi,K., König,M., Smith,L., Gu,S., Hellerstein,J., Sealfon,S.C. and Sauro,H.M. (2018) Tellurium notebooks—An environment for reproducible dynamical modeling in systems biology. *PLOS Comput. Biol.*, **14**, e1006220.

158. Saltelli,A., Annoni,P., Azzini,I., Campolongo,F., Ratto,M. and Tarantola,S. (2010) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput. Phys. Commun.*, **181**, 259–270.

159. Herman,J. and Usher,W. (2017) SALib : An open-source Python library for Sensitivity Analysis. *J. Open Source Softw.*, **2**, 3873–3878.

160. Saltelli,A. (2002) Making best use of model evaluations to compute sensitivity indices. *Comput. Phys. Commun.*, **145**, 280–297.

161. Lonard,D.M. and O'Malley,B.W. (2006) The Expanding Cosmos of Nuclear Receptor Coactivators. *Cell*, **125**, 411–414.

162. McDowell,I.C., Barrera,A., D'Ippolito,A.M., Vockley,C.M., Hong,L.K., Leichter,S.M., Bartelt,L.C., Majoros,W.H., Song,L., Safi,A., *et al.* (2018) Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome Res.*, **28**, 1272–1284.

163. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A., *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.*, **107**, 21931–21936.

164. Adomas,A.B., Grimm,S.A., Malone,C., Takaku,M., Sims,J.K. and Wade,P.A. (2014) Breast

tumor specific mutation in GATA3 affects physiological mechanisms regulating transcription factor turnover. *BMC Cancer*, **14**.

165. Takaku,M., Grimm,S.A., De Kumar,B., Bennett,B.D. and Wade,P.A. (2020) Cancer-specific mutation of GATA3 disrupts the transcriptional regulatory network governed by Estrogen Receptor alpha, FOXA1 and GATA3. *Nucleic Acids Res.*, **48**, 4756–4768.

166. Jiang,G., Wang,X., Sheng,D., Zhou,L., Liu,Y., Xu,C., Liu,S. and Zhang,J. (2019) Cooperativity of co-factor NR2F2 with pioneer factors GATA3, FOXA1 in promoting ERα function. *Theranostics*, **9**, 6501–6516.

167. Swaminathan,G., Nguyen,L.P., Namkoong,H., Pan,J., Haileselassie,Y., Patel,A., Ji,A.R., Mikhail,D.M., Dinh,T.T., Singh,H., *et al.* (2021) The aryl hydrocarbon receptor regulates expression of mucosal trafficking receptor GPR15. *Mucosal Immunol. 2021 144*, **14**, 852–861.

168. Öberg,M., Bergander,L., Håkansson,H., Rannug,U. and Rannug,A. (2005) Identification of the tryptophan photoproduct 6-formylindolo[3,2-b]carbazole, in cell culture medium, as a factor that controls the background aryl hydrocarbon receptor activity. *Toxicol. Sci.*, **85**, 935–943.

169. Pan,Y., Deng,M., Li,J., Du,B., Lan,S., Liang,X. and Zeng,L. (2020) Occurrence and Maternal Transfer of Multiple Bisphenols, including an Emerging Derivative with Unexpectedly High Concentrations, in the Human Maternal-Fetal-Placental Unit. *Environ. Sci. Technol.*, **54**, 3476–3486.

170. Andra,S.S., Austin,C., Yang,J., Patel,D. and Arora,M. (2016) Recent advances in simultaneous analysis of bisphenol A and its conjugates in human matrices: Exposure biomarker perspectives. *Sci. Total Environ.*, **572**, 770–781.

171. Liu,J., Li,J., Wu,Y., Zhao,Y., Luo,F., Li,S., Yang,L., Moez,E.K., Dinu,I. and Martin,J.W. (2017) Bisphenol A Metabolites and Bisphenol S in Paired Maternal and Cord Serum. *Environ. Sci. Technol.*, **51**, 2456–2463.

172. Boucher,J.G., Boudreau,A., Ahmed,S. and Atlas,E. (2015) In vitro effects of bisphenol A β-D-glucuronide (BPA-G) on adipogenesis in human and murine preadipocytes. *Environ. Health Perspect.*, **123**, 1287–1293.