UNIVERSITY OF CALIFORNIA, IRVINE
& SAN DIEGO STATE UNIVERSITY

Computational Models to Assess the Role of Environmental Exposures During Development

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computational Science

by

Ashley Valentina Schwartz

Dissertation Committee:
Associate Professor Uduak George (SDSU), Chair
Distinguished Professor Pierre Baldi (UCI), Co-Advisor
Associate Professor Karilyn Sant (SDSU), Co-Advisor
Chancellor's Professor John Lowengrub (UCI)
Assistant Professor Nicole Sparks (UCI)

2025

DEDICATION

To my loved ones.

"If we are going to live so intimately with these chemicals
eating and drinking them,
taking them into the very marrow of our bones -
we had better know something about their nature and their power."
- Rachel Carson, *Silent Spring*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply grateful to my dissertation committee for their invaluable guidance and unwavering support throughout this journey. I would like to thank to Dr. Uduak George for her steadfast mentorship and expertise in computational bio-mathematics, which have profoundly shaped both this dissertation and my development as a researcher. I am also sincerely thankful to Dr. Pierre Baldi for his support during my time at UCI and for introducing me to a vibrant research community in computer science. My gratitude also goes to Dr. Karilyn Sant for her continuous guidance in developmental toxicology and for providing a rich, dynamic laboratory environment where I could apply my computational skills. I appreciate Dr. John Lowengrub for his valuable feedback and support, and Dr. Nicole Sparks for sharing her expertise and insights in environmental toxicology.

I would especially like to express my profound gratitude to Dr. Uduak George and Dr. Karilyn Sant, whose joint mentorship has been instrumental in transforming me from a naive undergraduate student into the researcher I am today. Their shared passion for research and their dedication to fostering an interdisciplinary environment have opened doors to opportunities I never imagined possible, and for that I am extremely grateful.

I also wish to acknowledge the many members of the Computational Science Research Center (CSRC) community who have nurtured my academic and professional growth. My sincere thanks go to Dr. Jose Castillo for his leadership and commitment to creating opportunities for CSRC students, and to Dr. Satchi Venkataraman for his dedication to student professional development and guidance throughout the years. I am equally grateful to Parisa Plant and Jean Macneil for their many behind-the-scenes efforts that make the program so successful.

Finally, I would like to acknowledge those who have had a significant impact on my personal growth and success throughout my studies. To my family, for their unwavering love and support in all my endeavors and for encouraging me to always strive for my best; to Cameron Sacks, for his constant support and encouragement through every high and low; and to Isabel White, for walking through this journey along side me. I would not be where I am today without the love and support of all these remarkable individuals.

I have been fortunate to receive support through various scholarships, fellowships, and grants throughout the course of this research. I have personally been supported under the Association for Computing and Machinery Computational and Data Science Fellowship, the Academic Support & Scholarships for Interdisciplinary Computational Scientists under a National Science Foundation Grant (DUE-193046), the San Diego ARCS Scholar Award, the SDSU University Graduate Fellowship and the Computational Science Research Center. The research presented here has also been supported by the National Institute of Health (National Institute of Diabetes and Digestive and Kidney Diseases) (1R21DK134931-01), the National Institute of Environmental Health Sciences (K01ES031640), and a National Science Foundation Faculty Early Career Development (CAREER) award (DMS2240155).

The research presented in this dissertation comprises published works and materials committed for publication with co-authors. The text of Chapter 2 and Chapter 3 is an adaptation of the material as it appears in [1] and [2] respectively, published by the Oxford University Press with an Open Access Creative Commons Attribution License. The coauthors listed in these publications are Dr. Karilyn Sant and Dr. Uduak George who directed and supervised research which forms the basis for Chapter 2 and Chapter 3. The text of Chapter 4 has been submitted for publication. The coauthors listed in these publications are Arielle Beaulieu, who supported the data use agreement and data engineering, and Dr. Karilyn Sant and Dr. Uduak George who directed and supervised research which forms the basis for Chapter 4.

VITA

Ashley Valentina Schwartz

EDUCATION

Doctor of Philosophy in Computational Science **2025**
San Diego State University *San Diego, California*
University of California Irvine *Irvine, California*

Bachelor of Science in Applied Mathematics **2019**
San Diego State University *San Diego, California*

FELLOWSHIPS AND SCHOLARSHIPS

University of Graduate Fellow **2024–2025**
San Diego State University College of Graduate Studies

ARCS Scholar **2022–2025**
Achievement Rewards for College Scientists, San Diego Chapter

NSF S-STEM ASSICS Scholar **2020–2025**
National Science Foundation Funded Academic Support and Scholarships for Interdisciplinary Computational Scientists

ACM SIGHPC Computational and Data Science Fellow **2020–2022**
Association for Computing and Machinery Special Interest Group on High Performance Computing

HOWELL-CSUPERB Research Scholar **2018–2019**
Doris A. Howell Foundation – California State University Program for Education and Research in Biotechnology

RESEARCH EXPERIENCE

Graduate Research Fellow **2024–2025**
San Diego State University *San Diego, California*

Graduate Research Assistant **2019–2024**
San Diego State University *San Diego, California*

Quantitative Systems Pharmacology Intern **2022**
Takeda Pharmaceuticals *San Diego, California*

Graduate Research Fellow **2021–2022**
University of California, Irvine *Irvine, California*

Undergraduate Research Assistant **2018–2019**
San Diego State University *San Diego, California*

TEACHING EXPERIENCE

Python Programming Graduate Teaching Assistant **2021–2022**
University of California, Irvine *Irvine, California*

Calculus for Life Sciences Teaching Assistant **2020–2021**
San Diego State University *San Diego, California*

Precalculus Undergraduate Teaching Assistant **2017–2019**
San Diego State University *San Diego, California*

REFEREED JOURNAL PUBLICATIONS

Integrating network analysis and machine learning to elucidate chemical-induced pancreatic toxicity in zebrafish embryos **2025**
Toxicological Sciences

Catalytically distinct metabolic enzyme isocitrate dehydrogenase 1 mutants tune phenotype severity in tumor models **2025**
Journal of Biological Chemistry

danRerLib: a python package for zebrafish transcriptomics **2024**
Bioinformatics Advances

Development of a dynamic network model to identify temporal patterns of structural malformations in zebrafish embryos exposed to a model toxicant, tris(4-chlorophenyl)methanol **2023**
Journal of Xenobiotics

Spatial heterogeneity of excess lung fluid in cystic fibrosis: generalized, localized diffuse, and localized presentations **2022**
Applied Sciences

Mathematical modeling of the interaction between yolk utilization and fish growth in zebrafish, *Danio rerio* **2021**
Development

The ecotoxicological contaminant Tris(4-chlorophenyl) methanol (TCPMOH) impacts embryonic development in zebrafish (*Danio rerio*) **2021**
Aquatic Toxicology

Acetylcholine regulates pulmonary pathology during viral infection and recovery **2020**
Immunotargets and Therapy

The roles and benefits of using undergraduate student leaders to support the work of SUMMIT-P **2020**
Journal of Mathematics and Science: Collaborative Explorations

SUBMITTED JOURNAL PUBLICATIONS

Associations between birth defect incidence and maternal, sociodemographic, and environmental factors in California from 2018-2019: A computational approach **2025**
Submitted

Comparative toxicity of the fungicide Boscalid and its metabolite M510F01 in zebrafish embryos **2025**
Under Peer Review

SOFTWARE

danRerLib `https://sdsucomptox.github.io/danrerlib/index.html`
Python package to support zebrafish researchers through transcriptomics and functional enrichment analysis, improving on pathway identification for the model organism.

ABSTRACT OF THE DISSERTATION

Computational Models to Assess the Role of Environmental Exposures During Development

By

Ashley Valentina Schwartz

Doctor of Philosophy in Computational Science

University of California, Irvine, 2025

Associate Professor Uduak George (SDSU), Chair

Environmental exposures during critical periods of development pose significant public health challenges, yet traditional toxicological assessments often fall short in capturing the multi-faceted nature of developmental toxicity. In this dissertation, I present an integrative suite of computational models that harness diverse high-throughput data, from zebrafish transcriptomics to population-level birth records, to predict and elucidate the mechanisms underlying environmentally induced developmental toxicity.

At the molecular level, I developed a custom Python package, danRerLib, to analyze zebrafish transcriptomics data and improve developmental toxicity insights. Addressing limited zebrafish gene annotation, danRerLib uses orthology mapping to link zebrafish genes to human counterparts, enabling robust functional enrichment analysis with Gene Ontology and the Kyoto Encyclopedia of Genes and Genomes.

Expanding on these molecular insights, I applied network modeling and machine learning techniques to whole-embryo zebrafish transcriptomic datasets to extract organ-specific bioactivity profiles in the pancreas following exposure to a wide set of chemicals. This combined approach identified key driver genes and potential biomarkers of toxicity in the pancreas, demonstrating that whole-embryo datasets can be repurposed to gain insights into specific

organ responses to chemical exposures.

At the population level, I integrated environmental, demographic, and clinical data such as pregnancy, labor and delivery complications using multilayer network modeling and machine learning to unravel the complex interactions contributing to birth defect outcomes. By integrating these data layers, the models provide a more comprehensive understanding of how various risk factors, including environmental exposures and social determinants of health, interact to influence developmental toxicity.

Collectively, this work advances predictive toxicology by bridging molecular and population-level analyses to better understand environmentally induced developmental disorders. By leveraging network modeling and machine learning, these models identify key molecular drivers of developmental toxicity and uncover environmental and clinical factors contributing to birth defect risk.

Chapter 1

Introduction

Environmental exposures during critical periods of development are increasingly recognized as major determinants of lifelong health, yet understanding the complex interplay between exposure and adverse developmental outcomes remains a significant challenge. While traditional toxicological approaches have provided valuable insights, they often fall short in capturing the multifaceted nature of developmental toxicity, failing to fully account for the combined influence of chemical, biological, and social factors. This dissertation addresses this critical gap by developing a suite of integrative computational models that leverage diverse, high-throughput data sources, ranging from zebrafish transcriptomics to population-level birth records, to elucidate and predict the consequences of environmental exposures during development. Specifically, this work advances computational modeling strategies, integrating network modeling, multilayer network analysis, and machine learning, to build a predictive toxicological framework. This framework aims to not only identify key molecular mechanisms of toxicity in zebrafish but also assess the risk of birth defects in human populations, considering the influence of real-world environmental and social factors. This chapter provides a comprehensive introduction to the challenges facing the field of developmental toxicology, reviews the current landscape of modeling approaches, and outlines how

this research addresses existing gaps through targeted research aims.

1.1 Background

Developmental toxicity, adverse effects on the developing organism resulting from environmental exposures, is a critical public health concern. Environmental pollutants, such as pesticides, endocrine disruptors, and other emerging contaminants, have been linked to an increased risk of birth defects, reproductive disorders, and various other adverse health outcomes [3, 4, 5, 6, 7]. In severe cases, these abnormalities may result in preterm birth, spontaneous abortion, or infant mortality. Because developmental processes are highly sensitive to perturbation, even low-level exposures during critical periods can have lifelong consequences [8, 9, 10]. Because these issues may persistent or even manifest throughout the lifecourse, the economic and health burden on individuals can be quite high. Understanding the complex interplay between environmental exposures and developmental outcomes is crucial for effective prevention and intervention.

Studying the effects of environmental contaminants on human embryonic development is challenging due to both practical and ethical constraints. Historically, developmental toxicology has relied on model organisms—such as rodents or other mammals—to assess the impact of chemical exposures. However, there has been a strong desire to reduce the use of these models, which are hindered by longer gestational periods, higher costs, and difficulties in dynamically monitoring *in utero* development [11, 12]. While *in vitro* assays, such as those used in the EPA's ToxCast program, offer a higher-throughput approach, they struggle to predict complex *in vivo* outcomes, particularly in the context of developmental toxicity where multiple multiple cell networks, organ systems and developmental processes are involved [13].

To address these limitations, the field of developmental toxicology is increasingly leveraging alternative approaches, including public repositories for zebrafish and human population studies [14, 15]. Zebrafish (*Danio rerio*) have emerged as a valuable alternative model for developmental toxicology [16, 11]. Their genetic similarity to humans, rapid embryonic development, and transparent embryos that allow real-time observation of developmental processes make them ideal for high-throughput toxicological screening [17]. The availability of extensive zebrafish transcriptomic data, often deposited in public repositories like National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) database, provides a rich resource for investigating the molecular pathways perturbed by environmental contaminants [18]. However, interpreting zebrafish transcriptomic data remains a significant challenge. Incomplete pathway annotations and gaps in mechanistic understanding can hinder the full biological interpretation of these data. Moreover, while zebrafish offer a valuable model for studying molecular mechanisms, they cannot fully replicate the complex environmental and social factors that influence human development.

Complementing these experimental models, human population studies offer invaluable real-world data on exposure and health outcomes. Datasets from California state agencies, such as the California Department of Public Health's Birth Defects Monitoring Program and the California Office of Environmental Health Hazard Assessment's (OEHHA) CalEnviroScreen program, provide historical data on birth defects and environmental pollutants, respectively, incorporating information on the physical and social environment. These data offer the opportunity to investigate how factors such as socioeconomic status, environmental pollution levels, maternal pregnancy complications, and access to healthcare may contribute to the incidence of birth defects [19].

Despite the richness of these data sources, significant challenges remain in integrating them and realizing their full potential. Human population studies, while robust in real-world exposure information, are inherently complex. Multifactorial influences, including socioeconomic

status, demographic variability, and maternal health factors, can confound the analysis of environmental effects. To date, the analysis of these datasets has been limited to simplistic probabilistic or linear regression methods that are limited in their ability to capture the complex, multidimensional nature of these datasets and the interplay between health, environmental, and social factors [20, 21, 22, 23, 19]. This underscores the need for integrative and predictive computational approaches that can bridge the divide between controlled experimental models and the multifaceted reality of human health outcomes.

Recent advances in systems modeling, machine learning, and artificial intelligence offer powerful tools for dissecting the complex interplay of factors contributing to developmental toxicity [24, 25, 26, 27, 28, 29, 30]. However, several key challenges persist. For example, many functional enrichment analysis tools lack comprehensive gene annotations in zebrafish, which can hinder the full interpretation of transcriptomic data, potentially leading to biased or incomplete insights into the biological mechanisms affected by toxic exposures. Additionally, integrating whole-embryo datasets optimized for high-throughput toxicity testing into targeted-organ analyses remains difficult, as does synthesizing disparate data from multiple sources [31]. Addressing these limitations is critical for developing robust predictive models that can effectively inform environmental contaminant relationships to adverse health outcomes.

1.1.1 Research Goal

This dissertation is driven by three overarching research questions aimed at advancing our understanding of developmental toxicity through integrative computational modeling. First, can improved gene annotation pipelines for zebrafish, employing techniques like human orthology mapping, enhance our biological understanding of how environmental contaminants affect molecular pathways? Second, can publicly available whole-embryo zebrafish tran-

scriptomic data be leveraged to assess target-organ toxicity through an integrated modeling approach that combines network analysis and machine learning? Third, what environmental and social factors are most associated with birth defect outcomes, and can a novel modeling strategy that integrates disparate data sources elucidate these complex inter-relationships? In addressing these questions, this work develops scalable computational tools that bridge molecular insights with population-level analyses, ultimately advancing predictive toxicology.

1.2 Research Overview

This dissertation focuses on developing and utilizing computational models to understand how environmental contaminants affect embryonic development. Each computational approach and modeling strategy presented here is motivated by current gaps in the literature, with the goal of establishing a foundation for future work and creating a scalable resource as the volume of data in developmental toxicity continues to grow. The following chapters directly address the limitations identified in the previous section, representing three independent research projects that have culminated in three distinct research projects. While each chapter is freestanding, they build upon the computational approaches and insights developed in the preceding chapters, creating a cohesive and progressive narrative. An overview of the dissertation is provided in Figure 1.1, highlighting the connectivity and progression of each chapter.

Chapter 2 addresses the critical limitation of incomplete gene annotations in zebrafish, which hinders the accurate interpretation of transcriptomic data following exposure to environmental contaminants. Many functional enrichment analysis tools rely on comprehensive gene annotations, but zebrafish pathways are often incompletely annotated, potentially leading to biased or incomplete insights into the biological mechanisms affected by toxic exposures. To overcome this challenge, Chapter 2 introduces danRerLib, a novel Python package designed

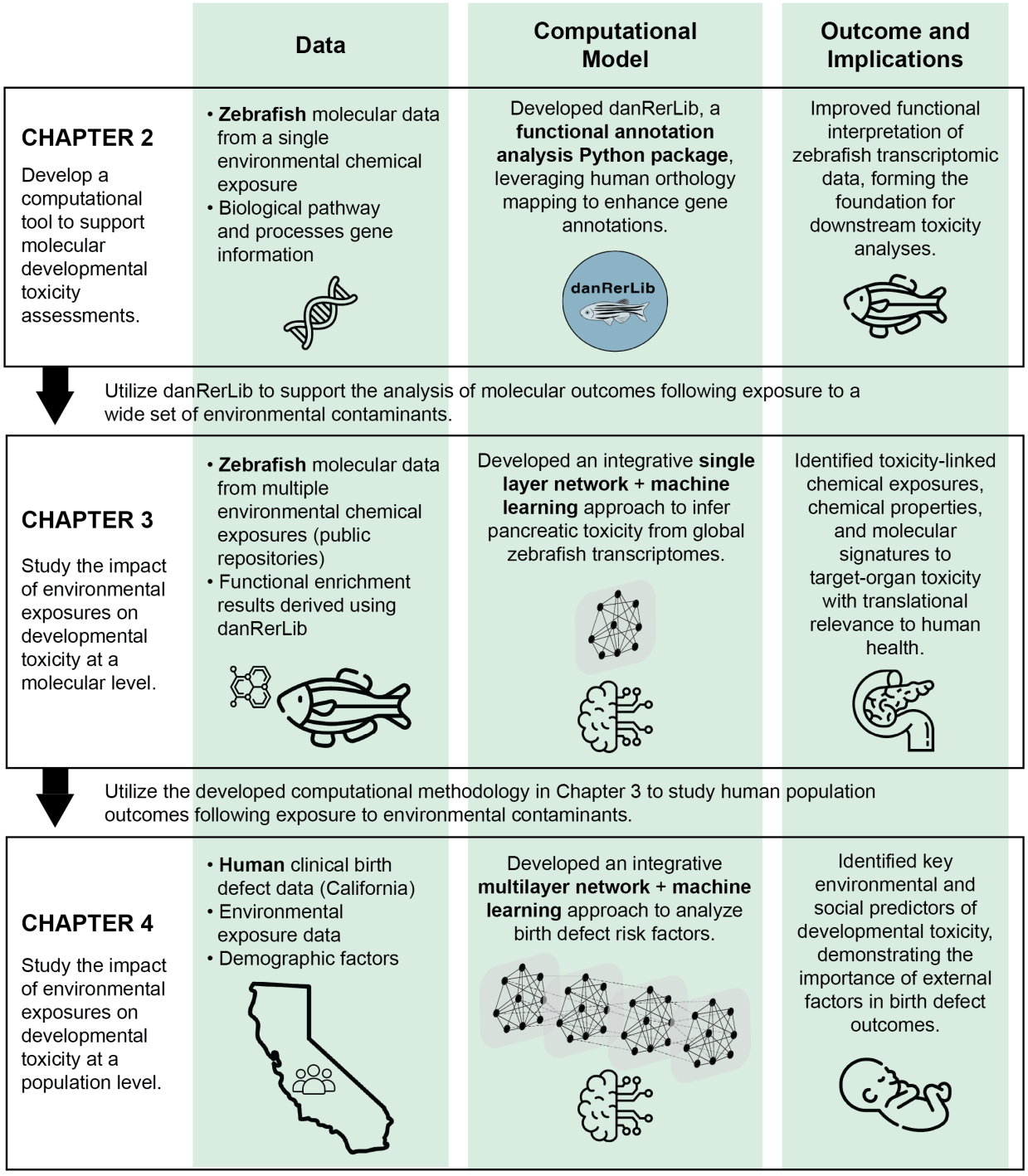

Figure 1.1: Overview of the dissertation.

to enhance functional enrichment analysis of zebrafish transcriptomic data. danRerLib leverages human orthology mapping to bridge the annotation gap, enabling researchers to access a more complete set of gene annotations and gain a more comprehensive understanding of the biological pathways impacted by environmental exposures. This chapter details the development and implementation of danRerLib, demonstrating its utility through a case study and comparative analyses.

Chapter 3 builds upon the functional enrichment capabilities developed in Chapter 2 and tackles the challenge of understanding molecular organ-specific developmental toxicity in zebrafish. Although whole-organism transcriptomic studies are optimized for high-throughput screening, they have traditionally been used to assess global effects of chemical exposures rather than the specific responses of individual organs. Chapter 3 addresses this limitation by developing an integrative computational framework that combines network analysis and machine learning to investigate organ-specific toxicity. Using pancreatic toxicity as a model, this chapter demonstrates how this framework can uncover relationships between chemical exposures and disrupted developmental pathways within a specific organ, identifying potential biomarkers and mechanistic insights. This work leverages publicly available zebrafish transcriptomic data, demonstrating the potential for *in silico* toxicology to reduce reliance on animal studies while still extracting valuable biological information.

Chapter 4 extends the computational framework developed in Chapter 3 to address the complex interplay of factors influencing birth defect outcomes in human populations. Human studies often analyze environmental exposures and clinical outcomes in isolation, neglecting the crucial interactions among health, environmental, and socioeconomic variables. Chapter 4 tackles this limitation by applying multilayer network modeling and machine learning to a large dataset of births in California. This approach allows for the integration of diverse data types, including environmental exposures, socioeconomic factors, and pregnancy, labor, and delivery complications, to investigate the complex interplay of factors influencing birth

defect outcomes. By constructing multilayer networks, this chapter aims to identify key predictors of birth defects and provide a scalable model for birth defect investigations.

1.3 Summary and Significance

In summary, this dissertation leverages advanced computational modeling techniques to address the complex challenge of understanding and predicting developmental toxicity resulting from environmental exposures. By integrating enhanced transcriptomic analysis, network modeling, and machine learning, this work provides a multi-scale framework that spans from molecular mechanisms in zebrafish to epidemiological outcomes in human populations. The innovative approaches presented here not only fill critical gaps in current modeling strategies but also offer scalable tools for future research in developmental toxicology and risk assessment. The findings of this study have significant implications for public health, potentially guiding interventions, regulatory policies, and the identification of emerging environmental threats.

Chapter 2

danRerLib: a Python Package for Zebrafish Transcriptomics

The motivation for Chapter 2 stems from my experiences as a zebrafish researcher confronted with the challenge of performing functional analysis in a species with incomplete pathway annotations. Existing bioinformatics tools, largely designed for more popular organisms such as rodents or humans, do not fully address the unique needs of zebrafish transcriptomics researchers. In response, we developed danRerLib, a Python package tailored specifically for zebrafish functional enrichment analysis. This tool not only underpins the research presented in this dissertation, including further investigations in Chapter 3, but also serves as a resource for the broader zebrafish research community. The work detailed in this chapter, which has been published in *Bioinformatics Advances* [1], outlines the design, implementation, and evaluation of danRerLib.

Understanding the pathways and biological processes underlying differential gene expression is fundamental for characterizing gene expression changes in response to an experimental condition. Zebrafish, with a transcriptome closely mirroring that of humans, are frequently

utilized as a model for human development and disease. However, a challenge arises due to the incomplete annotations of zebrafish pathways and biological processes, with more comprehensive annotations existing in humans. This incompleteness may result in biased functional enrichment findings and loss of knowledge. danRerLib, a versatile Python package for zebrafish transcriptomics researchers, overcomes this challenge and provides a suite of tools to be executed in Python including gene ID mapping, orthology mapping for the zebrafish and human taxonomy, and functional enrichment analysis utilizing the latest updated Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. danRerLib enables functional enrichment analysis for GO and KEGG pathways, even when they lack direct zebrafish annotations through the orthology of human-annotated functional annotations. This approach enables researchers to extend their analysis to a wider range of pathways, elucidating additional mechanisms of interest and greater insight into experimental results.

danRerLib, along with comprehensive documentation and tutorials, is freely available. The source code is available at `https://github.com/sdsucomptox/danrerlib/` with associated documentation and tutorials at `https://sdsucomptox.github.io/danrerlib/`. The package has been developed with Python 3.9 and is available for installation on the package management systems PIP (`https://pypi.org/project/danrerlib/`) and Conda (`https://anaconda.org/sdsu_comptox/danrerlib`) with additional installation instructions on the documentation website.

2.1 Introduction

Transcriptomics is a widely used methodology for assessing gene expression changes in response to experimental conditions and plays a pivotal role in understanding the molecular mechanisms governing biological processes and disease. RNA-sequencing is the most common

form of investigating differential gene expression changes that gives an unbiased snapshot of the transcriptome [32]. The zebrafish (*Danio rerio*) is an established vertebrate model for human development and disease, especially in embryonic toxicity studies and developmental biology [14, 33, 16]. With a transcriptome that closely mimics that of humans [34], zebrafish offer a unique opportunity to investigate the effects of experimental conditions on gene expression.

The quantification and interpretation of gene expression changes often involve performing functional enrichment analysis to discern the under- or over-expression of specific biological processes or pathways [35]. For example, pharmacologists and toxicologists commonly conduct functional enrichment analyses to assess the molecular impact of exposures or treatments. Through the analysis, researchers are able to identify key up and downregulated pathways following exposure [36, 37, 38]. However, existing tools frequently lack tailored support for zebrafish researchers, and updates to databases like Kyoto Encyclopedia of Genes and Genomes (KEGG) [39] or Gene Ontology (GO) [40] are not always readily available. An additional challenge is the fact that many pathways and biological processes, which are known to also occur in zebrafish, remain unannotated for zebrafish with more comprehensive annotations existing in humans. For instance, KEGG Release 109.1 (Mar 2024) lists 179 pathways for zebrafish and a total of 357 annotated for human. It is important to note that the KEGG Disease database is only designed for humans, which poses a challenge for researchers who want to utilize zebrafish as a model for relevant human diseases. The current limitations of existing tools may lead to incomplete, outdated, and biased functional annotation results, which in turn may hinder the interpretation of experimental outcomes.

To address these limitations, we introduce *Danio rerio* Library, danRerLib, a specialized Python package for zebrafish researchers utilizing transcriptomics in their research. Uniquely designed to support the latest genome and annotation builds for KEGG and GO, danRerLib extends conventional gene tools by providing dedicated gene conversion and functional en-

richment tools tailored explicitly for zebrafish studies. A primary contribution is the incorporation of orthology-based functional enrichment analysis, enabling the study and testing of pathways and biological processes not currently annotated for zebrafish.

2.2 Materials and Methods

The Python package danRerLib is divided into five key modules: the mapping module, the KEGG module, the GO module, enrichment module, and the enrichplots module. The functionality and purpose of each of the included modules are described below and highlighted in Figure 2.1A.

Mapping Module. The mapping module has been designed to convert Gene IDs from different databases and gene nomenclatures, including National Center for Biotechnology Information (NCBI) Entrez Gene IDs [41], Ensembl Gene IDs [42], Zebrafish Information Network (ZFIN) IDs [43], and official Gene Symbols [43]. In addition, the mapping module facilitates orthology mapping between human and zebrafish taxonomy, using orthologous genes defined and managed by ZFIN [43].

KEGG Module. The KEGG module contains tools for the investigation of genes within a KEGG pathway or KEGG disease. The KEGG pathway database contains pathway maps that encapsulate our knowledge of cellular and organism-level functions, categorized into metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases [39]. The KEGG disease database describes molecular networks caused by perturbants including gene variants, viruses and other pathogens, and various environmental factors [44]. One can download a list of genes given a KEGG pathway or disease ID in any supported Gene ID format.

GO Module. The GO module contains tools to probe information from the Gene Ontol-

ogy databases including GO Biological Processes (GO BP), GO Molecular Functions (GO MF), and GO Cellular Components (GO CC). Gene Ontology categorizes gene products into structured ontologies, providing valuable insights into their biological roles, molecular activities, and cellular compositions [45]. The GO module allows for the exploration of these ontologies, allowing users to query ontologies by GO ID and analyze genes based on their involvement in biological processes, molecular functions, and cellular components. One can download a list of genes given a GO ID in any supported Gene ID format.

Enrichment Module. The enrichment module is equipped with tools designed for conducting functional enrichment analysis of gene sets within KEGG pathways, KEGG diseases, and Gene Ontology (GO) ontologies. Users gain the capability to assess the degree of enrichment for gene sets based on their datasets of gene expression, determining the overrepresentation or underrepresentation of genes within specific gene sets. The enrichment module allows for flexibility to conduct a global analysis across all gene sets or focus on specific IDs or subsets. Users will receive a detailed data table containing all IDs of gene sets significantly enriched, depleted, upregulated, or downregulated depending on whether a directional test is chosen. Two methods of functional enrichment analysis are currently supported including overrepresentation analysis by Fisher's Exact Test and the logistic regression method [46, 47]. The p-value of Fisher's Exact Test is computed utilizing the hypergeometric distribution as shown in Equation 2.2.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{d}}{\binom{N}{a+c}} \tag{2.1}$$

In Equation (1), a is the number of genes in the gene set and significantly expressed, b is the number of genes not in the gene set that are significantly expressed, c is the number of genes in the gene set and nor significantly expressed, d is the number of genes neither in the gene set nor significantly expressed, and N is the total number of genes in the background.

The logistic regression model can be described using Equation 2.2.

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x \tag{2.2}$$

Where the explanatory variable x is defined as the negative log of the pvalues for differential expression, $-log$(p-value), p is the probability of a gene belonging to a gene set, α is the intercept and β is the slope parameter. The Wald test is then utilized to assess the evidence that β is significantly different from zero following a χ^2 distribution with one degree of freedom and is defined in Equation 2.3.

$$W = \left(\frac{\hat{\beta}}{s_{\hat{\beta}}}\right) \tag{2.3}$$

The slope parameter β corresponds to the log odds of belonging to a gene set and when $\beta > 0$ the gene set is considered to be enriched [47].

Enrichplots Module. The enrichplots module complements the enrichment module, providing users with powerful visualization tools to effectively interpret and communicate the results obtained from functional enrichment analyses. With the enrichplots module, users can create a variety of plots, including bar charts, volcano plots, and dot plots. Bar charts can show the distribution of enriched, depleted, upregulated, or downregulated gene sets across different categories. Volcano plots help users understand the relationship between statistical significance and fold change, offering a comprehensive view of significantly altered pathways. Dot plots offer an alternative visualization method, portraying key enrichment metrics such as the odds ratio, the significance of the enrichment of the pathway, and the percentage of significant genes within the pathway. A dot plot example is shown in Figure 2.1 with the remaining options illustrated in Figure A.1. These visualization tools enable users to extract meaningful insights and communicate the biological relevance of significantly altered gene sets more effectively.

The databases used for the described modules have been built using the latest genome builds and nomenclature from ZFIN (release Mar 18, 2024), NCBI (FTP release Mar 18 2024), KEGG, and GO at the time of development in March 2024 September 2023 from the respective databases. The KEGG data is accessed through web API to ensure the latest online data is being utilized (Release 109.1 API Mar 1, 2024). Database updates will be updated quarterly. If a user wishes to update to the latest database build, each database can be rebuilt using the building functions within the modules to ensure the latest information is being utilized.

2.3 Results

A primary contribution of danRerLib is the ability to streamline functional enrichment analysis for pathways and gene sets not annotated for zebrafish through orthology. To illustrate this contribution, we analyzed our previously published RNA-sequencing data of 4 days post fertilization whole embryo zebrafish following exposure to the environmental contaminant tris(4-chlorophenyl)methanol (TCPMOH) [37]. In [37], we performed functional enrichment analysis using LR Path [47], which relies solely on zebrafish annotations. In the current study, we performed functional enrichment analysis using danRerLib's cut-off free logistic regression method and compared the findings using zebrafish annotations alone to those incorporating additional pathways via orthology. This approach showcases the enhanced capabilities of danRerLib in expanding the scope of functional enrichment analysis beyond zebrafish annotated pathways.

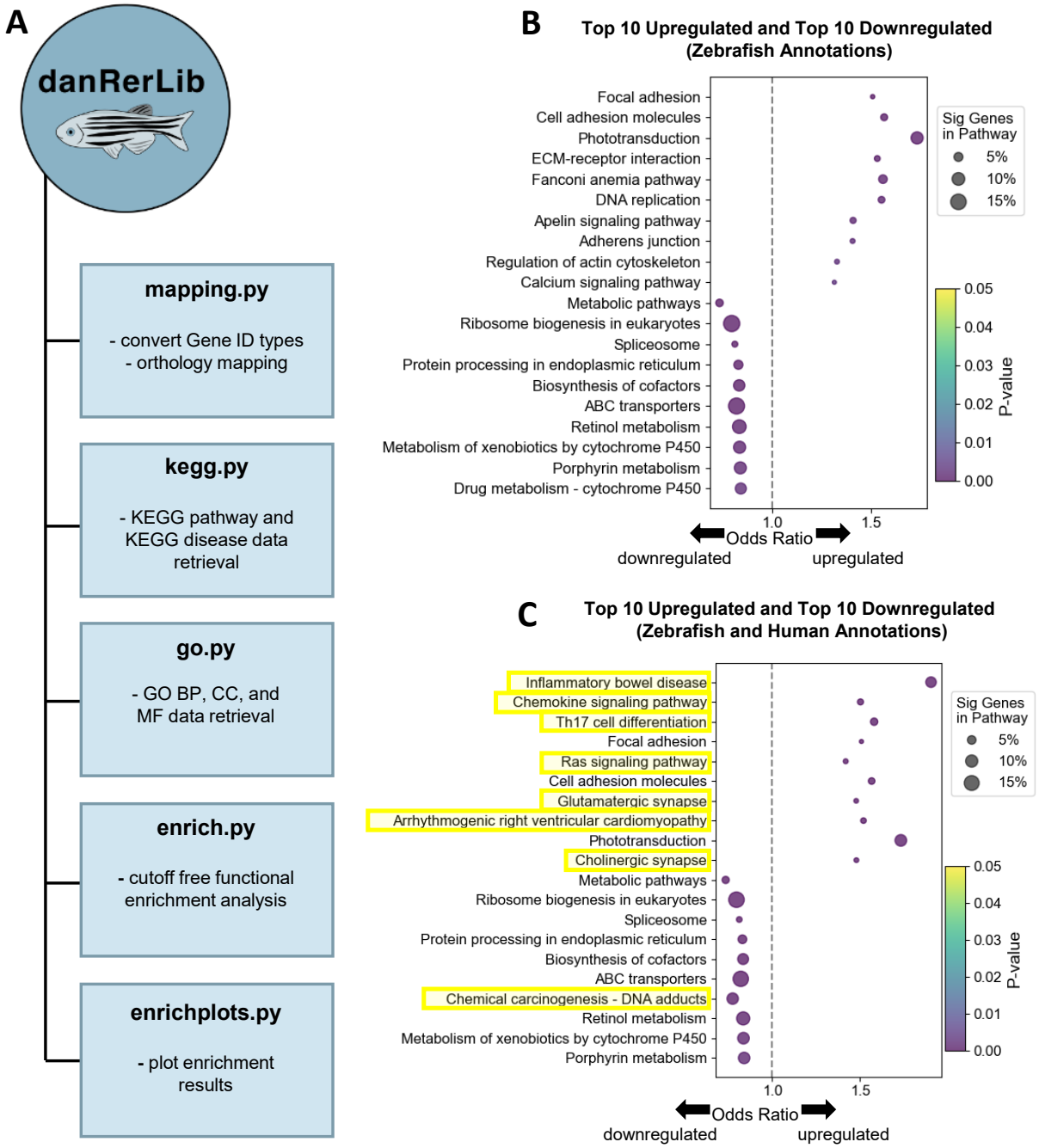

Figure 2.1: Overview of danRerLib modules, capabilities, and pathway enrichment results. (A) Diagram illustrating the core modules and key functionalities of the danRerLib software package designed for zebrafish pathway analysis. (B) Visualization of the top 10 upregulated and top 10 downregulated KEGG pathways identified using zebrafish-specific pathway annotations. (C) Visualization of the top 10 upregulated and top 10 downregulated KEGG pathways obtained by integrating zebrafish annotations with human pathway annotations via orthology. In panel (C), pathways that are uniquely discovered through orthology mapping are highlighted in yellow. In panels (B) and (C), the dot color indicates the statistical significance of the regulation (up or down), while the dot size represents the percentage of significant genes within each pathway.

2.3.1 Orthology-Based Enrichment Analysis and Expanded Pathway Identification

Our analysis revealed a significant increase in the number of identified KEGG pathways when incorporating orthology. We identified 55 significantly upregulated/downregulated KEGG pathways based on zebrafish annotations alone (p-value < 0.05). The top 10 up and top 10 downregulated pathways are shown in Figure 2.1B. After expanding the analysis to include orthology and implementing danRerLib's hybrid orthology enrichment, we identified a total of 95 KEGG pathways that were significantly upregulated or downregulated with the top 10 shown in Figure 2.1C. This is done by leveraging the 65% of human protein-coding genes having at least one zebrafish ortholog (based on the latest human genome v.GRC38), emphasizing the importance of orthology-based enrichment analysis in capturing biologically relevant information that may not be evident in zebrafish annotations alone. Out of the 55 pathways that were identified as significant using zebrafish annotations, 42 were also found to be present in the human annotation test. However, the identification of 13 significant pathways exclusively through zebrafish annotations unveils the complex relationship between human and zebrafish genomes and the need for a hybrid approach, leveraging zebrafish annotations when available.

2.3.2 Comparison of danRerLib to Existing Tools

To further evaluate danRerLib, we compared its functionalities and results to those of other prominent enrichment tools including DAVID [48, 49], gProfiler [50], FishEnrichr [51, 52], GSEApy [53], LRPath [47], clusterProfiler [54, 55], and GOAtools [56]. Table 2.1 outlines key differences, with danRerLib distinguished by three primary features: dedicated support for Python workflows in zebrafish transcriptomics, the use of a cut-off-free logistic regression method for enrichment analysis, and its hybrid orthology-based enrichment approach.

Among the tools listed, only four, gProfiler, GSEApy, GOATools, and danRerLib, offer Python support. Of these, danRerLib and gProfiler are the only tools regularly updated with the latest functional annotation builds for KEGG and GO. However, gProfiler lacks support for ZFIN gene IDs, requiring users to employ additional tools for zebrafish-specific analyses. In contrast, danRerLib natively supports ZFIN gene IDs and Python-based workflows.

Using all tools listed in Table 2.1 that support KEGG pathway analysis, we performed enrichment analysis and compared the total number of significant pathways identified at p-value < 0.05 (Table 2.2). danRerLib's orthology-based approach nearly doubled the number of tested pathways, from 179 to 357, substantially increasing the breadth of the analysis. Additionally, danRerLib supports pathway depletion analysis, a feature shared only by gProfiler and LR Path.

2.3.3 Comprehensive Documentation of danRerLib for Effective Use

In addition to its technical contributions, danRerLib's comprehensive and user-friendly documentation website plays a crucial role in facilitating its adoption by the research community 2.2. The documentation was crafted to ensure accessibility for users with varying levels of expertise in bioinformatics and zebrafish transcriptomics. It serves not only as a reference guide but also as an educational resource for understanding enrichment analysis and orthology-based methods.

The website includes detailed tutorials, use case examples, and an API reference that covers all core functions of danRerLib. Each tutorial walks users through key features such as data preparation, running enrichment analyses, and interpreting results. The examples provided are based on real-world datasets, including a replication of the results shown here, demonstrating how danRerLib can be applied to different experimental designs. The API reference

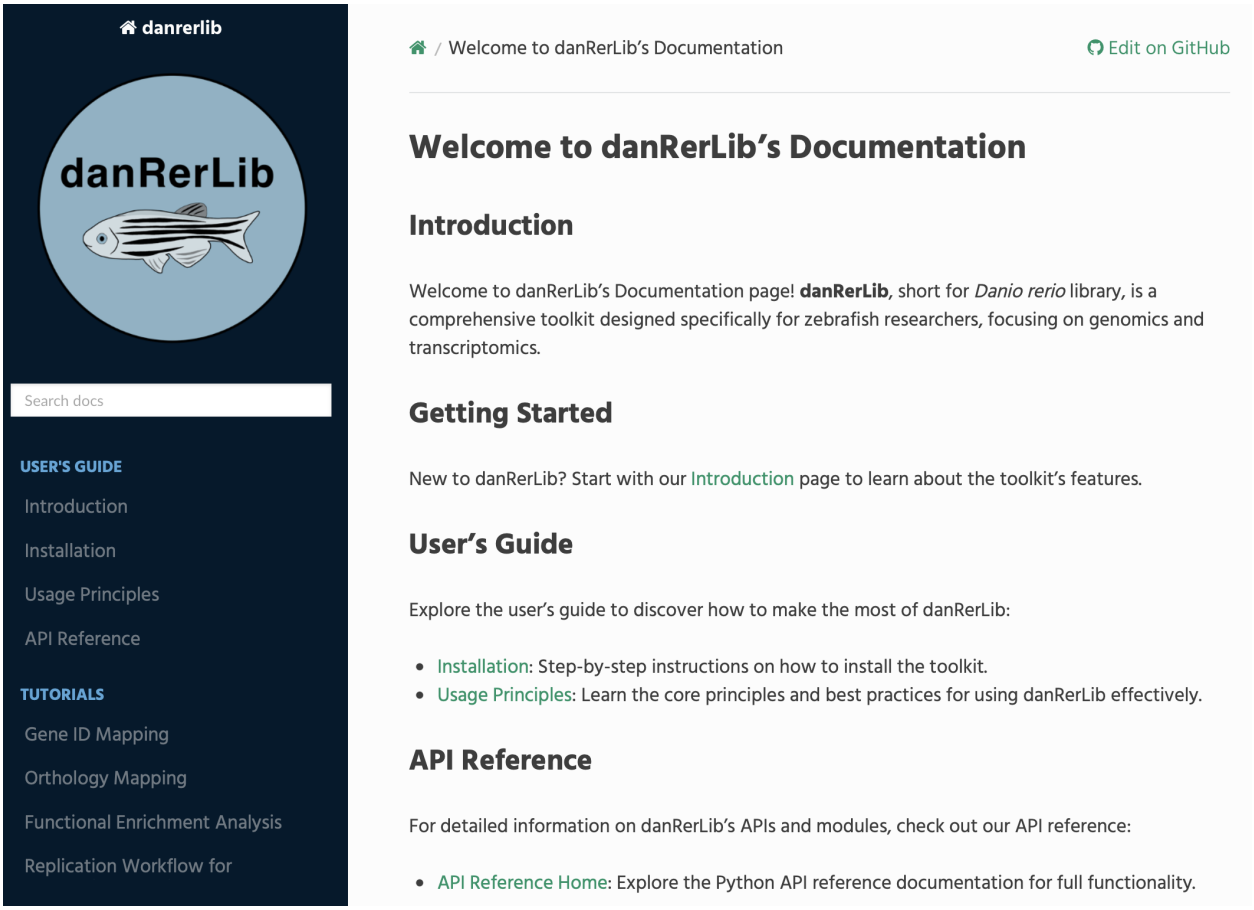

Figure 2.2: An illustration of the danRerLib documentation website, which provides comprehensive resources including installation instructions, tutorials, an API reference, and additional guides to assist users in implementing danRerLib for zebrafish pathway enrichment analyses. The webpage is accessible at: `https://sdsucomptox.github.io/danrerlib/index.html`.

section of the documentation offers a detailed breakdown of each function, including input requirements, output formats, and parameter descriptions. By providing this level of detail, users can confidently integrate danRerLib into their existing workflows without needing to decipher underlying code. Additionally, frequent updates to the documentation ensure that it remains aligned with the latest version of the package, reflecting any new features or improvements.

The website also highlights the community support facilitated by GitHub, where users can report issues, request new features, and engage with other researchers using danRerLib. This

collaborative approach fosters a growing community around zebrafish transcriptomics and promotes continuous improvement of the tool.

2.4 Discussion

danRerLib provides a comprehensive solution for functional enrichment analysis in zebrafish transcriptomics by expanding pathway coverage through orthology-based enrichment. Incorporating orthology led to the identification of nearly twice as many significant pathways compared to using zebrafish annotations alone. This finding underscores the importance of including cross-species annotations to capture a broader range of biological processes and potential disease mechanisms.

While KEGG lists 179 pathways for zebrafish, it includes 347 pathways for humans. By enabling users to test these additional human pathways, danRerLib fills a critical gap in zebrafish research. The ability to prioritize zebrafish-specific annotations while seamlessly integrating human pathways when necessary ensures that the results remain biologically relevant to the model organism. Unlike other tools that require manual mapping of zebrafish genes to human orthologs, danRerLib automates this process, streamlining the workflow for researchers.

One notable advantage of danRerLib over existing tools is its implementation of a cut-off-free logistic regression method developed by [47]. This method eliminates the need for arbitrary significance thresholds in defining differentially expressed genes, which can greatly influence enrichment results [57]. By offering this cut-off-free approach alongside traditional methods such as Fisher's exact test, danRerLib provides flexibility for researchers to choose the most appropriate method for their analysis. Additionally, because the logistic regression approach relies on up-to-date annotation databases, danRerLib ensures that the results remain current

and comprehensive.

The hybrid orthology approach introduced by danRerLib is particularly valuable in addressing the incomplete functional annotations available for zebrafish. The divergence in enrichment results between species is partly due to a whole genome duplication event in teleost fish, leading to the absence of one-to-one orthology for many genes. Based on the latest human genome build reported in NCBI (v.GRCh38) incorporated in the building of danRerLib, there are 20,644 human protein-coding genes and 13,508 of these genes have at least one zebrafish ortholog as reported in ZFIN; therefore, at least 65% of human genes have at least one zebrafish ortholog. To tackle the complexity of cross-species gene relationships, danRerLib offers a solution that allows users to selectively incorporate only human annotations that lack counterparts in zebrafish. This approach makes use of zebrafish annotations as the ground truth, enabling researchers to accurately interpret the significance of pathways in the context of zebrafish-specific genomic intricacies. By selectively incorporating human annotations only when zebrafish-specific counterparts are unavailable, danRerLib minimizes the risk of misinterpretation while maximizing pathway coverage. This approach allows researchers to explore human disease pathways that may otherwise be overlooked in zebrafish-only analyses.

One of the noteworthy aspects of the methodology proposed with danRerLib is the examination of pathways that are associated with human disease, given that KEGG disease annotations are available only for humans. Using zebrafish annotations alone as shown in Figure 2.1C, we discovered that KEGG metabolic pathways were considerably decreased. This result is further confirmed in Figure 2.1C with the inclusion of human annotations. However, by leveraging zebrafish annotations in KEGG and incorporating orthology, we uncovered additional insights. Notably, inflammatory bowel disease emerged as a top significantly upregulated pathway. This finding via danRerLib explains the pathological observation of intestinal effusion in the microscopy images of the zebrafish in [37]. Thus, danRerLib

provides valuable information about how TCPMOH may be linked to the underpinnings of disease. These findings using the streamlined orthology based enrichment analysis highlight the benefit of investigating both zebrafish and human annotations, a method often overlooked and requiring extra steps in existing methods. Introducing significant pathways and diseases can assist in comprehending complex disease manifestations that occur following exposure to chemicals.

Comparison with existing tools further highlights the unique capabilities of danRerLib. While several tools support KEGG pathway analysis, only danRerLib offers an orthology-based approach combined with Python-native workflows and regular annotation updates. Moreover, the inclusion of pathway depletion analysis ensures that researchers can comprehensively assess both upregulated and downregulated pathways, providing a balanced view of transcriptomic changes.

danRerLib provides a tailored and streamlined approach to gene ID conversions, orthology mapping, and functional enrichment analysis specifically for zebrafish researchers in Python. The incorporation of orthology enables researchers to overcome the limitations of direct zebrafish annotations, allowing for the exploration of pathways and biological processes not previously annotated. Though not all human diseases necessarily occur in zebrafish, nor are the same molecular processes necessarily conserved in disease progression between species, identification of these additional pathways can provide keen in-sights into pathophysiology. Given the growing popularity of Python among researchers, danRerLib provides a valuable resource for zebrafish researchers to complement bioinformatics workflows in Python while integrating multiple enrichment methodologies. Furthermore, danRerLib supports users of all computational experiences, providing comprehensive tutorials that guide users through the functionalities and maximize the package's utility.

Table 2.1: Comparative analysis of functional enrichment tools including danRerLib. Various enrichment analysis methods, including danRerLib, based on the statistical tests employed, frequency and scope of knowledgebase updates, and the extent of orthology support. Abbreviations used include: BH (Benjamini-Hochberg procedure for controlling the False Discovery Rate (FDR)), LR (Logistic Regression), and Fisher's (Fisher's Exact Test). This comparison provides insights into the strengths and limitations of each tool for pathway enrichment studies.

Enrichment Tool	**Statistical Test**	**Web-Based**	**Python Package**	**R Package**	**KEGG Knowledgebase Update**	**GO Knowledgebase Update**	**Adj p-value Method**	**Orthology Tool**	**Orthology Based Enrichment Analysis**	**Supports ZFIN IDs for Enrichment**
DAVID	Fishers	Yes	No	No	Updated November 2023	Updated May 2023	Bonferroni, BH FDR, Two-stage BH FDR	Yes	No	Yes
gProfiler	Fishers	Yes	Yes	Yes	Updated 01-22-2024	Updated January 2024	g:SCS threshold, Bonferroni, BH FDR	Yes	No	No
Fish-Enrichr	Fishers	Yes	No	Yes	KEGG 2019	GO 2018	BH FDR	Yes	No	No
GSEApy	Fishers	No	Yes	No	KEGG 2019	GO 2018	BH FDR	No	No	No
LRPath	LR	Yes	No	Yes	Bioconductor KEGG.db (2011)	Bioconductor KEGG.db (2011)	FDR	No	No	No
clusterProfiler	Fishers	No	No	Yes	Bioconductor KEGG.db (2011) or option for latest version using API	Bioconductor KEGG.db (2011)	BH FDR	Yes	No	No
GOAtools	Fishers	No	Yes	No	No KEGG support	Latest Version (updates automatically)	Bonferroni, BH FDR	No	No	No
danRerLib	Fishers and LR	No	Yes	No	Latest Version (updates automatically)	Latest Version (updates automatically)	Bonferroni, BH FDR	Yes	Yes	Yes

Table 2.2: Comparison of KEGG pathway functional enrichment analysis results for various enrichment methods including danRerLib. The number of included KEGG pathways is the number of zebrafish annotated pathways available to test against while the danRerLib orthology enrichment tool includes zebrafish annotated pathways when they exist and those mapped to zebrafish from human annotated pathways.

Enrichment Tool	**Number of Included KEGG Pathways**	**Number of Enriched KEGG Pathways**	**Number of Depleted KEGG Pathways**	**Number of Upregulated KEGG Pathways**	**Number of Downregulated KEGG Pathways**
DAVID	178	10	N/A	1	15
gProfiler	179	7	0	1	6
FishEnrichr	151	14	N/A	5	14
GSEApy	151	14	N/A	5	14
clusterProfiler	179	17	N/A	7	20
LRPath	155	17	12	25	13
danRerLib enrich_fishers	179	17	3	7	20
danRerLib enrich_logistic	179	16	17	29	26
danRerLib enrich_fishers orthology	357	26	14	17	27
danRerLib enrich_logistic orthology	357	24	50	60	35

Chapter 3

Integrating Network Analysis and Machine Learning to Elucidate Chemical-Induced Pancreatic Toxicity in Zebrafish Embryos

The motivation for Chapter 3 arose from the plethora of publicly available data gene expression data following exposure to a wide array of environmental chemicals. With an ever-increasing number of chemicals being identified, there is a unique need to support their toxicological evaluation using high-throughput methodologies. In particular, zebrafish whole-embryo datasets offer an attractive platform for high-throughput screening, and we hypothesized that these datasets could be further probed to extract organ-specific bioactivity insights. To this end, we sought to integrate network modeling with machine learning to predict the toxicity profile of chemicals on a target organ, specifically the pancreas, a well-known target of chemical-induced toxicity. By re-analyzing each included dataset through a uniform pipeline and applying danRerLib (developed in Chapter 2) for functional enrichment

testing, we aimed to combine mechanistic insights from transcriptomic data with chemical properties to predict organ-specific toxic effects. This approach not only lays the groundwork for investigating other target organs in the future but also capitalizes on the growing volume of zebrafish gene expression data. The results of this combined network and machine learning analysis, which is currently published in *Toxicological Sciences* [2], are presented in this chapter.

Zebrafish (*Danio rerio*) are a popular vertebrate model for high-throughput toxicity testing, serving as a model for embryonic development and disease etiology. However, standardized protocols using zebrafish tend to explore pathologies and behaviors at the organism level, rather than at the organ-specific level. This study investigates the effects of chemical exposures on pancreatic function in whole-embryo zebrafish by integrating network analysis and machine learning, leveraging widely-available datasets to probe an organ-specific effect. We compiled transcriptomics data for zebrafish exposed to 53 exposures from 25 unique chemicals, including halogenated organic compounds, pesticides/herbicides, endocrine-disrupting chemicals, pharmaceuticals, parabens, and solvents. All raw sequencing data were processed through a uniform bioinformatics pipeline for re-analysis and quality control, identifying differentially expressed genes and altered pathways related to pancreatic function and development. Clustering analysis revealed five distinct clusters of chemical exposures with similar impacts on pancreatic pathways with gene co-expression network analysis identifying key driver genes within these clusters, providing insights into potential biomarkers of chemical-induced pancreatic toxicity. Machine learning was utilized to identify chemical properties that influence pancreatic pathway response, including average mass, biodegradation half-life. The random forest model achieved robust performance (4-fold cross-validation accuracy: 74%) over XGBoost, support vector machine, and multiclass logistic regression. This integrative approach enhances our understanding of the relationships between chemical properties and biological responses in a target organ, supporting the use of zebrafish whole-embryos as a high-throughput vertebrate model. This computational workflow can be

leveraged to investigate the complex effects of other exposures on organ-specific development.

3.1 Introduction

Environmental pollutants and pharmaceuticals pose significant health risks, potentially contributing to various diseases and developmental disorders. With the continuous development of new chemical compounds, there is a growing need for effective toxicological evaluation. The EPA's Toxicity Forecaster (ToxCast) program has been instrumental in evaluating a diverse set of chemicals using *in vitro* assays and has included predictions for many non-evaluated chemicals using computational (*in silico*) models [58, 59, 60]. Though modernized approaches, including microphysiological systems and other new approach methodologies (NAMs) have aided in the accuracy of these predictive models for *in vivo* extrapolation, complex processes like embryonic development and other endocrine or multi-organ physiological effects can be difficult to predict *in vitro* [61, 13].

The zebrafish (Danio rerio) is a popular high-throughput vertebrate model for developmental toxicity studies, especially those utilizing non-targeted analysis for molecular and pathway bioactivity [11, 16]. Zebrafish are an advantageous model for developmental toxicology, offering large clutches of embryos, external fertilization, quick embryogenesis (3 days vs. 3 weeks in rodents) and similarity (>70%) to the human genome [34]. There is an ever-growing body of literature from our group and others that investigates the impacts of chemical exposures on embryonic development by assessing gene expression alterations following exposure using RNA sequencing (RNA-Seq). In a typical zebrafish toxicity study, whole embryos are pooled for RNA-Seq. This methodology is advantageous to support high-throughput screening, as it does not require the dissection of targeted organs—which is particularly challenging in zebrafish due to their small size. Additionally, exposures and embryo collection can be automated, minimizing human intervention and improving experimental efficiency [62].

However, one major limitation of this bulk RNA-Seq approach for zebrafish embryotoxicity studies is the inability to probe tissue- or organ-specific effects of a chemical. Because samples include a myriad of tissue and cell types at different states of differentiation, it can be difficult to make conclusions about the impacts of the exposure on the organogenesis of a specific tissue. In this way, zebrafish may yet be disadvantageous for organogenesis-stage studies compared to other mammalian models. However, computational approaches that exploit the known molecular pathways and responses of particular cells and tissues in zebrafish may help to overcome these limitations.

In this investigation, we utilize computational approaches to delve deeper into a wide set of whole embryo transcriptomic datasets to assess organ-specific gene expression, namely in the pancreas. The pancreas is a complex organ that governs metabolic functions and possesses both endocrine and exocrine functions. Pancreatic organogenesis is a highly conserved processes between vertebrates including zebrafish and humans [63] and has been shown to be particularly sensitive to chemical exposure [64, 65, 66, 67, 68, 69, 70]. In addition to structural similarities, pancreas disease states, such as diabetes, are conserved between zebrafish and humans [71, 72]. Given its importance and frequent involvement in toxicity responses, the pancreas provides a compelling case study for our computational framework.

To investigate genes involved in pancreatic development and function, we leverage the functional annotation databases Kyoto Encyclopedia of Genes and Genomes (KEGG) [39] and Gene Ontology (GO) Biological Processes (BP), Cellular Components (CC), and Molecular Functions (MF) [40]. By utilizing these classifications, we can efficiently analyze the vast amount of gene expression data generated from zebrafish exposed to a wide range of chemicals. Traditional computational analyses compare control and exposed groups to identify expression changes, using statistical techniques to identify specific genes and pathways affected by chemical exposure. However, these methods are limited for larger datasets.

To address this limitation, we propose a novel approach integrating network analysis and

machine learning. Gene co-expression networks have emerged as powerful methodologies for investigating complex gene expression correlations between chemical exposures [18, 73]. These networks, where nodes represent genes and links depict relationships, are particularly useful as gene interactions are crucial for understanding biological functions and disease. Network analysis, which can depict the complex interactions between entities, is widely used to investigate biological systems [74, 75]. These networks can reveal key driver genes involved in responses to chemical exposure. Additionally, supervised machine learning algorithms trained on labeled data can predict chemical toxicity based on chemical properties [76, 28].

To our knowledge, no study has yet integrated network analysis and machine learning to identify chemical properties influencing observed gene expression changes in zebrafish. We hypothesize such analysis can provide a comprehensive understanding of the relationship between chemical properties and the resulting biological response. Therefore, the primary goal of this study is to aggregate publicly available chemical exposure embryonic toxicity datasets and determine the relationships between chemical properties and pancreatic gene expression changes using network analysis and machine learning. This approach offers valuable insights into mechanisms underlying chemical-induced developmental toxicity and establishes a framework for future investigations utilizing zebrafish for comprehensive toxicological screening.

3.2 Materials and Methods

3.2.1 Data Acquisition and Study Inclusion Criteria

Transcriptomic datasets from zebrafish embryos exposed to environmental pollutants and pharmaceuticals were obtained from the Gene Expression Omnibus (GEO) [77], a publicly accessible repository maintained by the US National Center for Biotechnology Information

(NCBI) that hosts a diverse collection of high-throughput sequencing datasets. To ensure dataset relevance and comparability across studies for downstream analysis, certain criteria were established. Inclusion criteria encompassed (1) the availability of raw transcriptomic RNA sequencing data in GEO, (2) RNA-Seq sourced from pooled whole zebrafish embryos between 4- and 7-days post-fertilization (dpf), (3) exposure to environmental stimuli, (4) exposure occurred during development, and (5) exposure administered through the aquatic environment. Studies that did not meet these inclusion criteria were excluded from further analysis. To ensure a systematic selection process, we applied both automated and manual exclusion steps. Automated exclusion removed studies that did not contain relevant keywords related to development (e.g. development, larval), or exposure (e.g. chemical, exposed, treated). Manual exclusion was performed by reviewing study meta-data and full-text and removing those that sequenced isolated tissues instead of whole embryos, collected embryos outside the 4-7 dpf window, or used an invalid exposure type, such parental or dietary exposures, as well as those examining chemical mixtures rather than individual chemical effects. A detailed breakdown of the selection process is provided in Figure 3.1. From the 21 datasets meeting inclusion criteria in GEO, we retrieved 26 unique chemicals and 53 distinct chemical exposures, as shown in Table 3.1. Here, "unique chemicals" refer to individual chemical compounds included in the dataset, whereas "distinct chemical exposures" account for multiple experimental conditions per chemical, such as different concentrations and exposure durations. Each chemical exposure was categorized into one of the following groups: Halogenated Organic Compounds (HOCs), Pesticides/Herbicides, Endocrine Disrupting Chemicals, Pharmaceuticals/Drugs, Parabens, and Solvents/Others.

Upon identification of relevant datasets meeting the criteria, a thorough screening process was conducted to ensure data quality and consistency for subsequent differential expression analysis. This screening process aimed to mitigate potential biases arising from technical variability and study-specific factors which have been shown to impact meta-analysis of zebrafish transcriptomic studies [78].

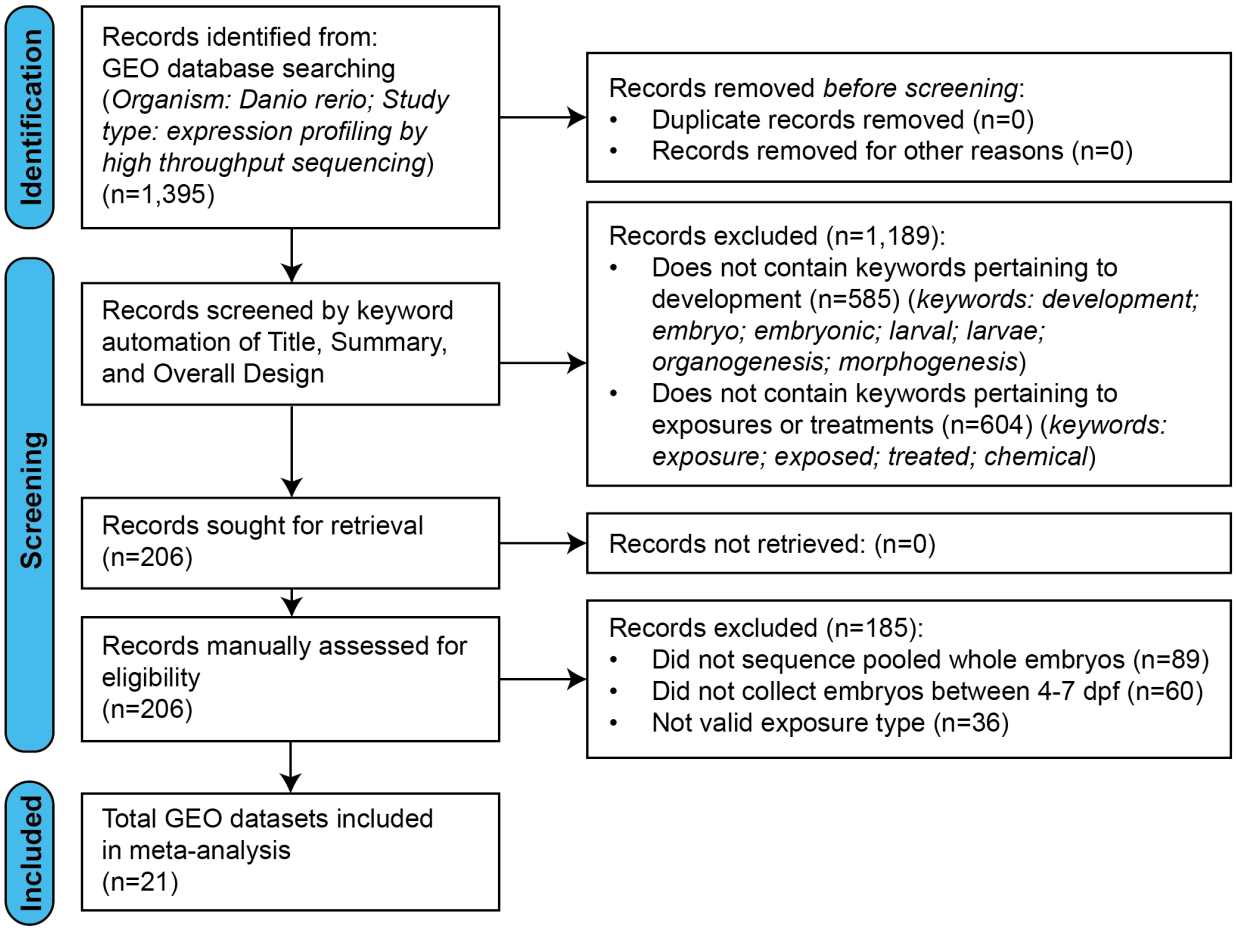

Figure 3.1: PRISMA flow diagram of GEO dataset screening and selection. The systematic process used to screen and select datasets from the Gene Expression Omnibus (GEO). After applying predefined inclusion and exclusion criteria, a final set of 21 GEO datasets was retained, representing 26 unique chemicals and 53 distinct chemical exposures.

Chemical properties for each chemical exposure were obtained from the Environmental Protection Agency CompTox Dashboard [79, 80]. Properties retrieved from the dashboard include the average mass, monoisotopic mass, physiochemical properties, and environmental fate and transport. The full list of chemical properties is provided in Tables B.1 and B.2. For each chemical property, the reported predicted median was used for chemical property analysis.

Table 3.1: Summary of chemical exposures curated from the Gene Expression Omnibus (GEO). The dataset comprising of 53 distinct chemical exposures from 21 GEO datasets, representing 26 unique chemicals and 53 distinct chemical exposures included in the dataset. Each exposure was selected based on predefined criteria and categorized into chemical groups such as Halogenated Organic Compounds, Pesticides/Herbicides, Endocrine Disrupting Chemicals, Pharmaceuticals/Drugs, Parabens, and Solvents/Others.

GSE	**Chemical Full Name**	**Chemical Short Name**	**Chemical Label(s)**	**Dose(s)**	**Chemical Type**	**Age (dpf)**	**Citation**
GSE114356	Perfluorobutanesulfonic Acid	PFBS	PFBS_16uM, PFBS_32uM	16 μM, 32 μM	Perfluoroalkyl Substances	4	[68]
GSE118955	3,3'-Dichlorobiphenyl	PCB-11	PCB11_20.2uM	20.2 μM	Halogenated Organic Compounds	4	[81]
GSE144525	Perfluorooctanesulfonic acid	PFOS	PFOS_0.6uM_A, PFOS_4.1uM	0.6 μM, 4.1 μM	Perfluoroalkyl Substances	4	[82]
GSE164074	Perfluorooctanesulfonic acid	PFOS	PFOS_40uM	40 μM	Perfluoroalkyl Substances	4	[83]
GSE164074	Sodium p-perfluorous nonenoxybenzene sulfonate	OBS	OBS_32uM, OBS_48uM	32 μM, 48 μM	Perfluoroalkyl Substances	4	[83]
GSE164128	Cannabidiol	CBD	CBD_0.5uM	0.5 μM	Pharmaceuticals/ Drugs	4	[84]
GSE164128	Tetrahydrocannabinol	THC	THC_4uM	4 μM	Pharmaceuticals/ Drugs	4	[84]
GSE165920	Tris(4-chlorophenyl) methane	TCPM	TCPM_50nM _acute	50 nM (acute)	Halogenated Organic Compounds	4	[37]

Continued on next page

Table 3.1 – continued from previous page

GSE	Chemical Full Name	Chemical Short Name	Chemical Label(s)	Dose(s)	Chemical Type	Age (dpf)	Citation
GSE165920	Tris(4-chlorophenyl) methanol	TCPMOH	TCPMOH_50nM_acute	50 nM (acute)	Halogenated Organic Compounds	4	[37]
GSE198106	Tris(4-chlorophenyl) methane	TCPM	TCPM_50nM	50 nM	Halogenated Organic Compounds	4	[70]
GSE198106	Tris(4-chlorophenyl) methanol	TCPMOH	TCPMOH_50nM	50 nM	Halogenated Organic Compounds	4	[70]
GSE232317	Boscalid	Boscalid	Boscalid_8.7nM	8.7 nM	Pesticide/ Herbicide	4	GEO
GSE232317	M510F01	M510F01	M510F01_8.4nM	8.4 nM	Pesticide/ Herbicide	4	GEO
GSE113676	Bisphenol A	BPA	BPA_0.4uM, BPA_4.4uM, BPA_17.5uM	0.4 μM, 4.4 μM, 17.5 μM	Endocrine Disrupting Chemicals	5	[85]
GSE125072	Perfluorooctanesulfonic acid	PFOS	PFOS_60nM, PFOS_0.6uM_B, PFOS_2uM	60 nM, 0.6 μM, 2 μM	Perfluoroalkyl Substances	5	[86]
GSE139599	Tributyltin	TBT	TBT_3nM, TBT_30nM, TBT_100nM	3 nM, 30 nM, 100 nM	Endocrine Disrupting Chemicals	5	[87]
GSE146199	Triptolide	Triptolide	Triptolide_1.6uM	1.6 μM	Pharmaceuticals/ Drugs	5	[88]
GSE239773	Ethyl-Parahydroxybenzoates	EtP	EtP_1.5uM, EtP_7.3uM, EtP_72.6uM	1.5 μM, 7.3 μM, 76.2 μM	Paraben	5	[89]
GSE239773	Methyl-Parahydroxybenzoates	MtP	MtP_3.1uM, MtP_15.4uM, MtP_154.1uM	3.1 μM, 15.4 μM, 154.1 μM	Paraben	5	[89]
GSE239773	Propyl-Parahydroxybenzoates	PrP	PrP_0.49uM, PrP_2.42uM, PrP_24.2uM	0.5 μM, 2.4 μM, 24.2 μM	Paraben	5	[89]
GSE80221	Cortisol	Cortisol	Cortisol_1uM	1 μM	Pharmaceuticals/ Drugs	5	[88]
GSE93310	2,2',4,4',5,5'-hexachlorobiphenyl	PCB-153	PCB153_0.1uM, PCB153_1uM, PCB153_10uM	0.1 μM, 1 μM, 10 μM	Halogenated Organic Compounds	5	[90]
GSE98741	3,3',4,4',5-Pentachlorobiphenyl	PCB-126	PCB126_0.3nM, PCB126_1.2nM	0.3 nM, 1.2 nM	Halogenated Organic Compounds	5	[91]
GSE168194	Tritiated water	HTO	HTO_3.3uM	3.3 μM	Solvents/Other	7	[92]
GSE172111	Ethanol	EtOH	EtOH_217mM	217 mM	Solvents/Other	7	[93]

Continued on next page

Table 3.1 – continued from previous page

GSE	Chemical Full Name	Chemical Short Name	Chemical Label(s)	Dose(s)	Chemical Type	Age (dpf)	Citation
GSE217875	Citalopram	CIT	CIT_30.8nM, CIT_3.1uM	30.8 nM, 3.1 μM	Pharmaceuticals/ Drugs	7	[94]
GSE217875	Ifosfamide	IFO	IFO_3.8nM, IFO_0.4uM	3.8 nM, 0.4 μM	Pesticide/ Herbicide	7	[94]
GSE224522	Afidopyropen	Afid	Afid_0.2nM, Afid_1.7nM	0.2 nM, 1.7 nM	Pesticide/ Herbicide	7	GEO
GSE224522	Atorvastatin	AVS	AVS_0.2nM, AVS_1.8nM	0.2 nM, 1.8 nM	Pharmaceuticals/ Drugs	7	GEO
GSE236484	Broflanilide	Broflanilide	Broflanilide _1.5nM, Broflanilide _15.1nM	1.5 nM, 15.1 nM	Pesticide/ Herbicide	7	[95]

3.2.2 RNA-Seq Quality Control, Alignment, and Differential Expression Analysis

Raw RNA-Seq data obtained from each study underwent standardized processing through a uniform bioinformatics pipeline to mitigate potential bias stemming from software and data analysis variability across studies (Figure 3.2, Step 1). This added preprocessing step effectively minimizes experimental and technical variations across studies and was automated using an in-house bash script. Initially, raw FASTQ files were retrieved from the Sequence Read Archive (SRA) run selector and subjected to quality assessment using FASTQC v0.12.1 [96]. Reads were assessed for quality scores of 20 or above, and any below-threshold reads were removed through trimming using cutadapt v4.4 [97]. The reference genome index was built using STAR with the NCBI zebrafish genome GRCz11 and annotation file. Subsequently, alignment to the zebrafish genome was performed using STAR v2.7.10b, followed by quantification of gene expression levels for each sample within the dataset using featureCounts v 2.0.1 [98, 99]. Given the scale of data involved, all computational analyses were

executed on a high-performance computing (HPC) platform consisting of 14 nodes, each with 16-20 cores and up to 264 GB of RAM per node. This HPC platform enabled efficient management of the extensive dataset and facilitated the deployment of an automated bash script for parallelizing the software utilized across compute nodes.

To determine differential expression patterns between treatment conditions and respective controls, the DESeq2 package was employed to calculate log2 fold changes in expression [100]. Each chemical sample was compared to its corresponding control within the original study to account for any inherent study design biases. All gene differential expression and associated p-values for each chemical exposure were aggregated into a consolidated data file for comprehensive analysis. Following aggregation, the dataset comprised 22,889 genes with log2 fold change values and associated p-values for each chemical exposure, indicative of a balanced gene expression dataset despite stemming from multiple studies.

3.2.3 Curation of Pancreas-Related Pathways and Genes

We manually curated a list of pancreas-related genes, pathways and diseases available through gene and pathway annotation databases in order to ascertain impacts of chemical exposures on the developing pancreas from a whole-embryo sample (Figure 3.2, Step 2). Pathways, biological processes, and disease annotations related to pancreas function and development were curated from two primary sources: the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway and Disease database, and the Gene Ontology (GO) database encompassing Biological Processes (BP), Molecular Functions (MF), and Cellular Components (CC) [39, 40]. We obtained a total of 100 curated gene sets spanning pancreatic functions and disease such as development, glucoregulation, cancer, and diabetes. For simplicity, each gene set is referred to as pathways regardless of whether the gene set was obtained from KEGG Pathway, KEGG Disease, GO BP, GO CC, or GO MF. A comprehensive list of curated

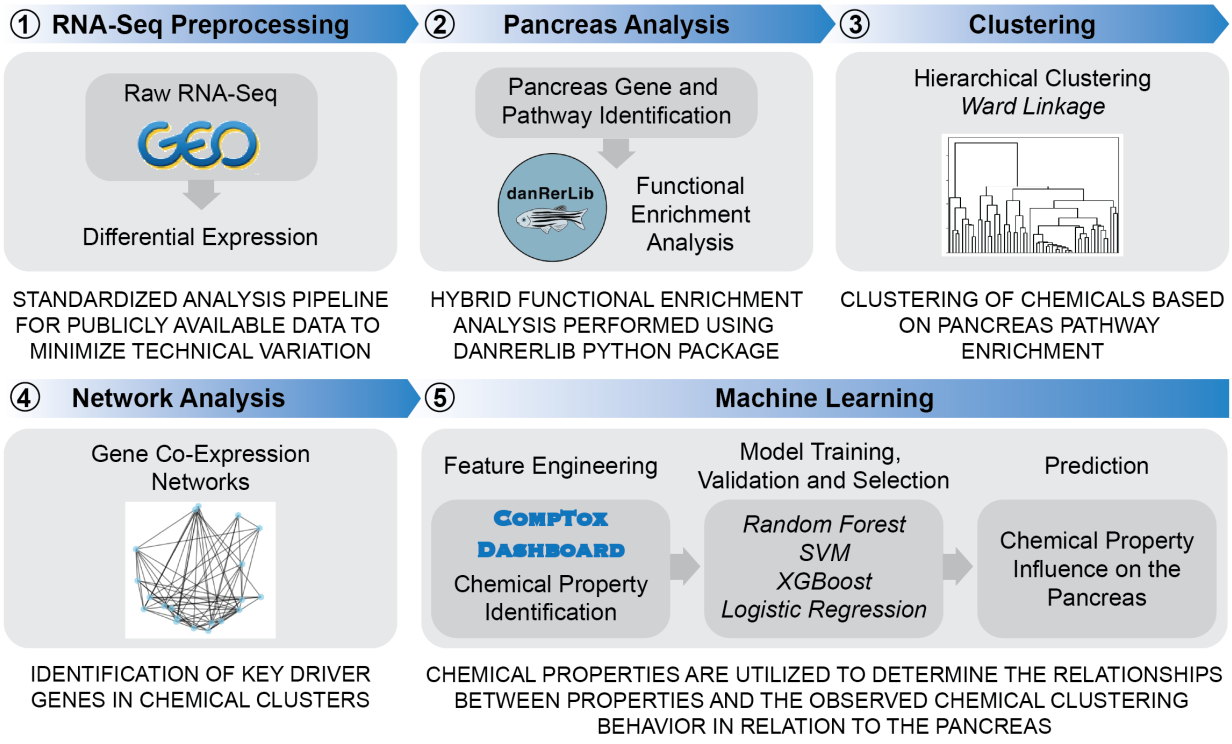

Figure 3.2: Schematic overview of the analysis pipeline. Following data curation and selection, this flowchart details the subsequent analysis steps: data preprocessing, feature extraction, machine learning model training, and model validation. The pipeline also integrates pathway enrichment and network analyses to predict cluster membership for chemical exposures, offering a clear and reproducible workflow.

pathways is provided in Table S2 and categorized into four groups: pancreas development, diabetes, pancreas diseases (including pancreatic cancer), and glucoregulation.

The genes related to pancreas function are those included within at least one of the curated pancreas pathways. After identifying the pancreas-related genes, a total of 736 genes with log2 fold change values and associated p-values were detected for each chemical exposure. This focused dataset of 736 genes provides a targeted framework for analyzing the impact of chemical exposures on pancreas-related genes, offering valuable insights into the differences in gene expression between chemical exposures.

3.2.4 Pancreas Pathway and Disease Functional Enrichment Analysis

To identify biological processes and disease mechanisms potentially affected by the chemical exposures, we performed pathway enrichment analysis on the curated pancreas-related pathways from KEGG and GO (Figure 3.2, Step 2). However, we found many of these pathways lacked complete gene annotations for zebrafish, while having complete annotations for humans. To overcome this limitation and ensure comprehensive analysis, we utilized danRerLib (v0.1.5), a specialized Python package designed for zebrafish functional enrichment studies [1]. This tool facilitates the examination of pathways lacking zebrafish-specific annotations by leveraging human orthology. Each chemical exposure underwent evaluation for enrichment across all curated pancreas-related pathways. When zebrafish-specific annotations were available, they were prioritized; otherwise, human-annotated pathways were tested via orthology, ensuring comprehensive pathway analysis despite potential species-specific annotation limitations (Table B.3). Enrichment analysis results were quantified using log odds ratios and associated p-values for each pathway of interest computed from the cut-off free logistic regression method facilitated by danRerLib [46, 1].

3.2.5 Clustering of Chemicals Based on Pancreatic Pathway Response

To identify patterns and relationships between different chemicals based on their impact on pancreatic function, unsupervised machine learning in the form of hierarchical clustering was employed (Figure 3.2, Step 3). This method allowed for the visualization of the similarity and divergence of pathway responses to various chemical exposures. Clustering was performed using the log odds ratio of enrichment for pancreas-related pathways. More specifically, hierarchical agglomerative clustering with Ward linkage, which minimizes the

within-cluster variance, was applied to the matrix of log odds ratio of enrichment using the SciPy Python package v1.13.0 [101]. This algorithm iteratively merged clusters of chemicals with the highest similarity in their pathway enrichment profiles, producing a dendrogram that illustrates the hierarchical relationships between the chemicals. By clustering chemicals based on the log odds ratio of pathway enrichment, distinct groups of chemicals with similar impacts on pancreatic function were identified. This methodological approach facilitated the understanding of shared and unique impacts on pancreatic function among various environmental exposures.

3.2.6 Pancreas Gene Co-expression Network Construction and Centrality Analysis

Pancreas-specific gene co-expression networks were utilized within these clusters to pinpoint key driver genes potentially contributing to the observed effects (Figure 3.2, Step 4). These gene co-expression networks were developed in Python using the NetworkX library v3.3 [102]. Pancreas genes are defined as those involved in at least one of the curated pancreas pathways as described previously (Table B.3). To identify the key pancreas genes that drive network connections, two types of networks were created: one encompassing all chemical exposures and one for each distinct cluster of chemicals identified in the clustering analysis.

For each chemical exposure, the log2 fold change for the 736 pancreas-related genes were extracted as described previously. Pearson correlation coefficients were computed for all pairs of genes to measure the strength of their co-expression. Separate correlation matrices were generated for the entire dataset of chemical exposures and for each chemical cluster. The Pearson correlation coefficient, r, between two genes, i and j, is calculated using the

formula in Eq 3.1.

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j})}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x_i})^2 \sum_{k=1}^{n} (x_{jk} - \bar{x_j})^2}} \tag{3.1}$$

In Eq 3.1, x_{ik} and x_{jk} are the log2 fold change values for genes i and j for chemical exposure k, and $\bar{x_i}$ and $\bar{x_j}$ are the mean log2 fold change values of genes i and j across all exposures or within a cluster.

Next, we converted the correlation matrices into undirected weighted gene co-expression networks. Each gene co-expression network is represented by a mathematical graph $G = (N,\ E,\ f : N \times N)$ where N is the set of nodes, E is the set of edges, and f is the mapping function that links node pairs through edges [103]. Each node $n_i \in N$ in the network represents a pancreas gene and each edge $e_{i,j} \in E$ represents the significant correlations between any gene pairs n_i and n_j. For an edge to be included within the network, Pearson's correlation between any two genes must exceed the predefined threshold $|r| > 0.7$, ensuring network edges represent strong correlations. The graph mapping function $f : N \times N$ may be represented by an adjacency matrix, A, that indicates the weighted edges composing the network topology. Specifically, for entries $a_{ij} \in A, a_{ij} = |r_{ij}|$ if $|r| > 0.7$ and 0 otherwise.

Network centrality analysis was conducted on the networks to identify key pancreas genes governing network topological properties. Key pancreas genes investigated include the genes with the top degree centrality and the top betweenness centrality [104]. A high degree centrality indicates a gene that has many connections within the network and is therefore highly correlated with other genes. Degree centrality D for node n_i in the weighted co-expression network is computed as $D(n_i) = \sum_{j \neq i} A_{ij}$. A high betweenness centrality indicates a gene that has a key position as a link between larger clusters of co-expressed genes. Betweenness centrality $B(n_i)$ for node n_i is computed as $B(n_i) = \sum_{s \neq n_i \neq t} \frac{\sigma_{st}(n_i)}{\sigma_{st}}$ where σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(n_i)$ is the number of those paths

that pass through node n_i. The nodes, or genes, that have high degree centrality and/or betweenness centrality are referred to as "hubs" and "bottlenecks" respectively.

3.2.7 Machine Learning Prediction of Chemical Effects on the Developing Pancreas

To better understand the chemical properties that influence altered gene expression, we employed supervised machine learning techniques (Figure 3.2, Step 5). Chemical properties were aggregated from the CompTox Dashboard to create a comprehensive chemical property feature space (Tables B.1 and B.2). Additionally, experimentally relevant features were incorporated, including micromolar concentration, zebrafish age at RNA collection, exposure duration, repeated exposure (static vs. repeated dosing), and fish strain. Missing values within the feature space were handled using median imputation [105] to ensure the completeness and robustness of the dataset. The target variable for the machine learning model was the cluster identifier for each chemical exposure. The primary objective was to predict the Cluster ID of a chemical, and consequently infer its impact on pancreatic function, based chemical properties while accounting for biological and experimental properties rather than direct gene expression alterations.

Several machine learning algorithms were tested, including random forest, XGBoost classifier, support vector machine (SVM), and multiclass logistic regression. To ensure robust and reliable performance metrics that are not dependent on a single train/test split, the models were trained and validated using a stratified 4-fold cross-validation approach. This approach helps maintain the relative distribution of classes in each fold, thereby preserving the original class proportions throughout the cross-validation process [106].

Feature importance was assessed to identify which chemical properties were most influential in determining the cluster assignment using SHAP (SHapley Additive exPlanations) [107].

SHAP values are a method to explain the output of machine learning model using model features and comes from the SHAP values for cooperative game theory. The formula for SHAP values can be explained by Eq 3.2, where K is the set of all features, S is a subset of the features excluding i, $f_x(S \cup \{i\})$ represents the instance x with only the features in subset S and including feature i, and $f_x(S)$ is the prediction of the model when only the features in S are known [108].

$$\phi\left(x_i\right) = \sum_{S \subseteq Ki} \frac{|S|!\left(K - |S| - 1\right)!}{K!}\left[f_x\left(S \cup \{i\}\right) - f_x\left(S\right)\right] \tag{3.2}$$

Multiple machine learning validation metrics were computed, including accuracy, precision, recall, and the F1-score [109]. Accuracy measures the proportion of times the model correctly predicts Cluster ID. Precision measures the proportion of true positives among all positive predictions. Recall measures the proportion of true positives among all actual positives. The F1-score is the harmonic mean between both precision and recall. Precision, recall, and F1-score were weighted according to class membership to account for any imbalances in the dataset. The mean and standard deviation of each metric across all cross validations were computed. Machine learning model building, and validation was conducted using the SciPy Python package [101].

3.3 Results

3.3.1 Unveiling Chemical Clusters Based on Pancreatic Function Alterations

To determine the similarities and differences in pancreatic function responses to various chemical exposures, hierarchical clustering was employed on the log odds ratio of pathway

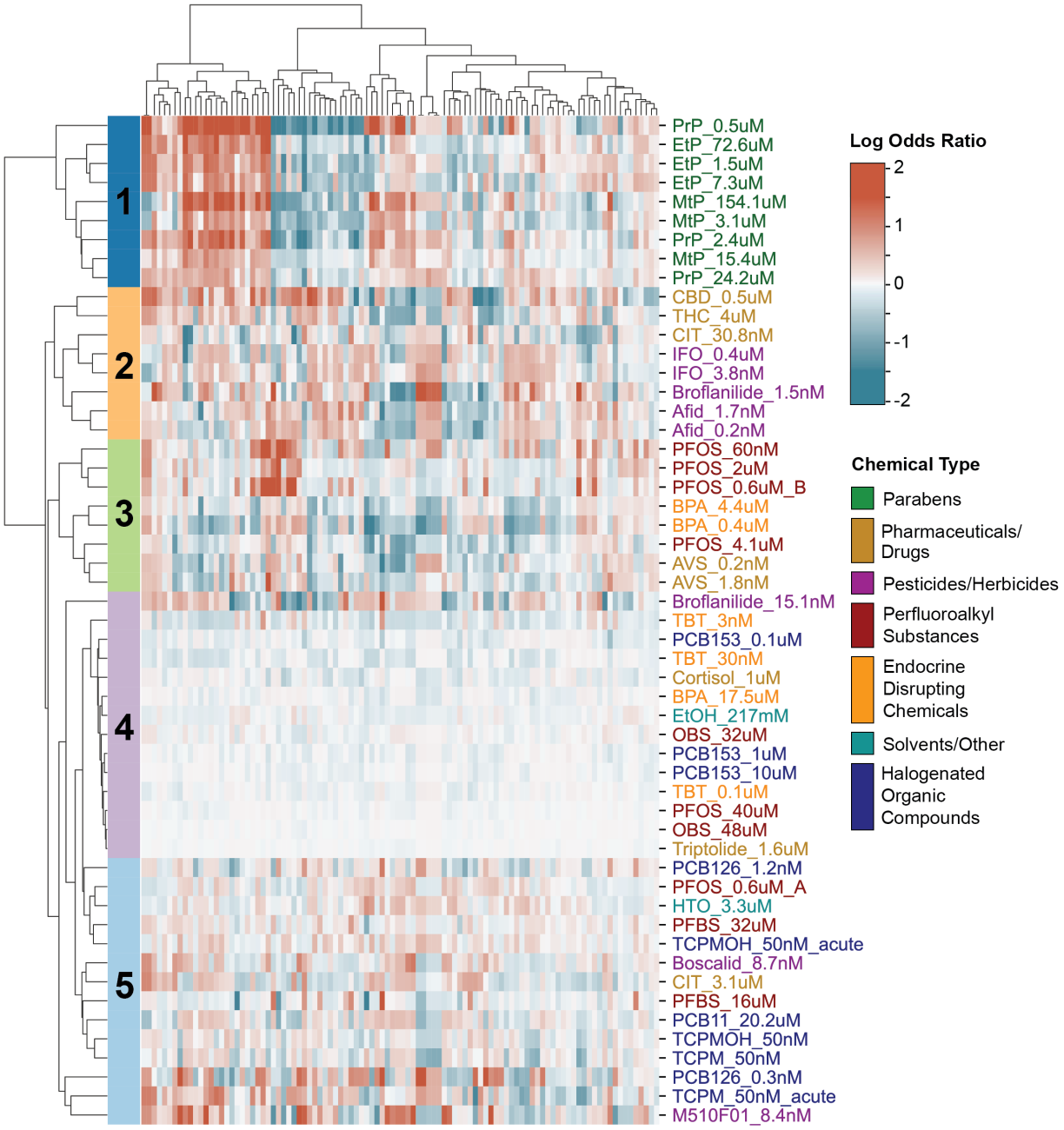

Figure 3.3: Heatmap of zebrafish pancreas pathway enrichment response to 53 chemical exposures. The log odds of enrichment for 100 curated pathways related to pancreatic function. Darker red shades indicate upregulation while darker blue shades denote downregulation. Each chemical exposure is annotated by its chemical type and assigned to one of five clusters (labeled 1–5) shown along the left.

enrichment scores for each chemical. This analysis allowed for the identification of distinct clusters of chemicals with similar impacts on pancreatic pathways, revealing patterns of

pathway response across different chemicals. Hierarchical clustering resulted in five distinct clusters, as illustrated in Figure 3.3. Clusters 4 and 5 were identified by visually subdividing a larger cluster, which included both Clusters 4 and 5, despite not strictly adhering to the dendrogram cutoff.

Cluster 1 represented all nine paraben chemical exposures included within the dataset and was the most distinct from any other chemical exposure responses. This suggests that the pancreas pathway response to parabens is unique among the chemical types, regardless of exposure level. Cluster 2 comprised of the majority of pesticide/herbicides chemical exposures and three other pharmaceutical drug chemical exposures. Cluster 3 included perfluoroalkyl substances (PFAS), other endocrine-disrupting chemicals, and the pharmaceutical atorvastatin (AVS). Interestingly, four of the six perfluorooctanesulfonic acid (PFOS) exposures were found in Cluster 3, with the others found in Clusters 4 and 5. Clusters 4 and 5 represented all of the HOC exposures with Cluster 4 containing the remaining endocrine disrupting chemicals and Cluster 5 containing no endocrine disrupting chemicals that are not classified as PFAS.

3.3.2 Distinct Cluster Effects on Pancreatic Function Revealed by Pathway Enrichment and Gene Expression

To better understand the differences in cluster impact on pancreas function, we examined the top 5 average up and downregulated pancreas pathways (Table 3.2) and the top up and downregulated pancreas genes (Table 3.3). This analysis revealed clear distinctions in the impact of each cluster on pancreatic function.

Cluster 1, which is composed of paraben chemical exposures, is characterized by dysregulation of insulin secretion, suggesting potential direct effects on the insulin-producing pancreatic β cells. The top upregulated pathways in this cluster include the negative regulation

of insulin secretion and insulin secretion involved in cellular response to glucose stimulus, among other pathways involved in glucoregulation. The clustering outcome suggests that paraben exposures are characterized by alterations to pathways and genes related to insulin production and secretion. Among the top upregulated genes in this cluster are *pkma* (pyruvate kinase M1/2a) and *pfkpa* (phosphofructokinase, platelet a). The *pkma* and *pfkpa* genes are a member of the glycolysis pathway and their upregulation indicates increased glycolytic activity [110]. Additional upregulated genes include *cpa1* (carboxypeptidase A1), *irs2b* (insulin receptor substrate 2b), and *ptprna* (protein tyrosine phosphatase receptor type Na). The gene *cpa1* encodes for a well-studied digestive enzyme shown to be critical to exocrine pancreas development [111] while *irs2b* is found in the insulin receptor binding and insulin receptor signaling pathway. The gene *ptprna* belongs to the protein tyrosine phosphatase (PTP) gene family, which plays a critical role in cell signaling and may influence insulin receptor signaling pathways [112].

Cluster 2, which comprises the majority of pesticide/herbicide chemical exposures and three pharmaceutical drug chemical exposures, exhibits significant alterations in pathways related to glucagon signaling and insulin sensitivity. This cluster exhibits upregulation in pathways related to glucagon family peptide binding and negative regulation of insulin receptor signaling, suggesting a state of altered insulin sensitivity. Among the top upregulated genes in this cluster are *slc2a8* (solute carrier family 2 member 8) and *stat3* (signal transducer and activator of transcription 3). The *slc2a8* gene is predicted to be involved in glucose transport [113]. The *stat3* gene is a key component to the insulin resistance pathway and a variety of disease pathways including pancreatic cancer. Additionally *stat3* activation has been found to be directly related to insulin resistance and sensitivity [114]. Other notable upregulated genes include *mlxip* (MLX interacting protein) and *cacna1sa* (calcium voltage-gated channel subunit alpha1 S). The *mlxip* gene has been shown to play a role in insulin resistance in humans [115], while *cacna1sa* is crucial for calcium ion transport, which is essential for insulin secretion and pancreatic function [116].

Cluster 3, which includes chemicals such as PFAS and endocrine-disrupting chemicals like Bisphenol A (BPA), along with the pharmaceutical atorvastatin, is characterized by significant alterations in pathways related to pancreatic development and disease. The top upregulated and downregulated pathway in this cluster include the negative regulation of canonical Wnt signaling and branching involved in pancreatic morphogenesis respectively. Other downregulated pathways include pancreatic δ cell differentiation and exocrine pancreatic insufficiency. Among the top upregulated genes in this cluster are *dpp4* (dipeptidyl peptidase 4), *tm4sf4* (transmembrane 4 L six family member 4), and *atp1a2a* (ATPase, Na+/K+ transporting, alpha 2a polypeptide). The *dpp4* gene is involved in glucose metabolism and human orthologs of this gene are implicated in type 2 diabetes mellitus [117]. The gene *tm4sf4* is involved in several negative relation pathways related to pancreas development including pancreatic A and B cell differentiation. Key downregulated genes in this cluster include *cyp26a1* (cytochrome P450, family 26, subfamily a, polypeptide 1) and *isl2b* (ISL LIM homeobox 2b). The gene *cyp26a1* is involved in retinoic acid metabolism, a crucial signaling pathway for cellular differentiation and development [118, 119, 120]. The downregulation of *isl2b*, a transcription factor, suggests possible exocrine pancreas development impacts.

Cluster 4 contains a number of well-known endocrine disrupting compounds such as BPA, cortisol (a hormone), PCB-153 (non-coplanar), and tributyltin. Cluster 4 is characterized by significant disruptions in pancreatic development, glucoregulation, and pancreatic disease pathways. The top upregulated pathways include those related to exocrine pancreatic insufficiency and dyserythropoietic anemia, tropical calcific pancreatitis, insulin secretion involved in the cellular response to glucose stimulus, and pancreatic A cell differentiation and fate commitment. These findings indicate that these chemicals heavily influence both the development and the functional aspects of pancreatic cells. Among the top upregulated genes in this cluster are *hspd1* (heat shock 60 protein 1) and *pygl* (phosphorylase, glycogen, liver). Both *hspd1* and *pygl* are involved in glucoregulation with *hspd1* associated with glu-

cose intolerance and *pygl* associated with glycogen catabolic processes and glycogen storage disease.

Cluster 5, predominantly influenced by exposures to HOCs, namely several coplanar PCBs (PCB-11, PCB-126) and the DDT-related compounds tris(4-chlorophenyl)methane and -methanol (TCPM and TCPMOH). Cluster 5 exhibits alterations in pathways related to glucoregulation and endocrine disruption in zebrafish. The top upregulated pathways in this cluster include cellular responses to insulin and glucagon stimuli, glucose transmembrane transport, and glucagon processing, indicative of an alteration to pancreatic function and glucose metabolism. Among the significantly upregulated genes in this cluster are *calm3a* (calmodulin 3a), *pparaa* (peroxisome proliferator-activated receptor alpha a), and *rpl3* (ribosomal protein L3), implicated in calcium signaling, digestive system, and exocrine pancreas development, respectively. Conversely, several downregulated pathways in Cluster 5, including those related to insulin secretion and canonical Wnt signaling, suggest suppressive effects on essential metabolic processes.

Table 3.2: Top average upregulated and downregulated pathways for each cluster. The top five average upregulated and downregulated pancreas-related pathways for each of the five identified clusters as defined by the log odds ratio of pathway enrichment. Each pathway is classified into one of the following categories: pancreas development, diabetes, pancreatic diseases, or glucoregulation.

	Pathway Type	**Pathway ID**	**Pathway Name**	**Pathway Category**	**Average Log Odds Ratio**
Cluster 1					
Up	GO BP	GO:0046676	negative regulation of insulin secretion	Glucoregulation	1.369
	GO BP	GO:0035773	insulin secretion involved in cellular response to glucose stimulus	Glucoregulation	1.321
	GO BP	GO:0038016	insulin receptor internalization	Glucoregulation	1.319
	GO BP	GO:2000078	positive regulation of type B pancreatic cell development	Development	1.276
	GO BP	GO:1901143	insulin catabolic process	Glucoregulation	1.245
Down	KEGG Disease	H00932	Tropical calcific pancreatitis	Pancreatic Disease	-1.306

Continued on next page

Table 3.2 – continued from previous page

	Pathway Type	Pathway ID	Pathway Name	Pathway Category	Average Log Odds Ratio
	KEGG Disease	H00933	Hereditary pancreatitis; Hereditary chronic pancreatitis	Pancreatic Disease	-1.118
	GO BP	GO:2000081	positive regulation of canonical Wnt receptor signaling pathway involved in controlling pancreatic B cell proliferation	Development	-0.934
	GO BP	GO:0044342	type B pancreatic cell proliferation	Development	-0.835
	KEGG Disease	H00543	Renal-hepatic-pancreatic dysplasia	Development	-0.829
Cluster 2					
Up	GO MF	GO:0120022	glucagon family peptide binding	Glucoregulation	0.666
	GO BP	GO:0046627	negative regulation of insulin receptor signaling pathway	Glucoregulation	0.617
	GO BP	GO:0003311	pancreatic D cell differentiation	Development	0.553
	KEGG Disease	H01228	Insulin-resistant diabetes mellitus with acanthosis nigricans; Type A insulin resistance	Diabetes	0.531
	KEGG Disease	H00942	Rabson-Mendenhall syndrome	Diabetes	0.531
Down	GO BP	GO:1900076	regulation of cellular response to insulin stimulus	Glucoregulation	-1.018
	GO BP	GO:0003327	type B pancreatic cell fate commitment	Development	-0.739
	GO BP	GO:0003326	pancreatic A cell fate commitment	Development	-0.683
	GO BP	GO:0003329	pancreatic PP cell fate commitment	Development	-0.683
	GO BP	GO:0046676	negative regulation of insulin secretion	Glucoregulation	-0.671
Cluster 3					
Up	GO BP	GO:2000080	negative regulation of canonical Wnt signaling pathway involved in controlling type B pancreatic cell proliferation	Development	1.037
	KEGG Disease	H02198	Pancreatic agenesis and congenital heart disease	Development	0.986
	GO BP	GO:1900078	positive regulation of cellular response to insulin stimulus	Glucoregulation	0.72
	GO MF	GO:0004967	glucagon receptor activity	Glucoregulation	0.623
	KEGG Disease	H00933	Hereditary pancreatitis; Hereditary chronic pancreatitis	Pancreatic Disease	0.581
Down	GO BP	GO:0061114	branching involved in pancreas morphogenesis	Development	-0.689
	KEGG Disease	H00920	Exocrine pancreatic insufficiency, dyserythropoietic anemia, and calvarial hyperostosis	Digestion	-0.655
	GO BP	GO:0003311	pancreatic D cell differentiation	Development	-0.641

Continued on next page

Table 3.2 – continued from previous page

	Pathway Type	Pathway ID	Pathway Name	Pathway Category	Average Log Odds Ratio
	GO BP	GO:0061178	regulation of insulin secretion involved in cellular response to glucose stimulus	Glucoregulation	-0.634
	KEGG Disease	H00513	Transient neonatal diabetes mellitus	Diabetes	-0.631
Cluster 4					
Up	KEGG Disease	H00920	Exocrine pancreatic insufficiency, dyserythropoietic anemia, and calvarial hyperostosis	Digestion	0.628
	KEGG Disease	H00932	Tropical calcific pancreatitis	Pancreatic Disease	0.485
	GO BP	GO:0035773	insulin secretion involved in cellular response to glucose stimulus	Glucoregulation	0.388
	GO BP	GO:0003310	pancreatic A cell differentiation	Development	0.386
	GO BP	GO:0003326	pancreatic A cell fate commitment	Development	0.361
Down	KEGG Disease	H00854	Wolfram syndrome; DIDMOAD syndrome	Diabetes	-0.356
	GO BP	GO:1901143	insulin catabolic process	Glucoregulation	-0.317
	GO BP	GO:2000081	positive regulation of canonical Wnt receptor signaling pathway involved in controlling pancreatic B cell proliferation	Development	-0.317
	GO BP	GO:0071377	cellular response to glucagon stimulus	Glucoregulation	-0.299
	GO MF	GO:0004967	glucagon receptor activity	Glucoregulation	-0.251
Cluster 5					
Up	KEGG Disease	H00920	Exocrine pancreatic insufficiency, dyserythropoietic anemia, and calvarial hyperostosis	Digestion	0.149
	GO BP	GO:0046628	positive regulation of insulin receptor signaling pathway	Glucoregulation	0.083
	GO BP	GO:0032869	cellular response to insulin stimulus	Glucoregulation	0.076
	GO BP	GO:1904659	glucose transmembrane transport	Glucoregulation	0.075
	KEGG Disease	H00019	Pancreatic cancer	Pancreatic Disease	0.069
Down	GO BP	GO:0035774	positive regulation of insulin secretion involved in cellular response to glucose stimulus	Glucoregulation	-0.182
	GO BP	GO:2000081	positive regulation of canonical Wnt receptor signaling pathway involved in controlling pancreatic B cell proliferation	Development	-0.176
	GO MF	GO:0120022	glucagon family peptide binding	Glucoregulation	-0.167
	GO BP	GO:0120116	glucagon processing	Glucoregulation	-0.163

Continued on next page

Table 3.2 – continued from previous page

	Pathway Type	Pathway ID	Pathway Name	Pathway Category	Average Log Odds Ratio
	GO BP	GO:0070092	regulation of glucagon secretion	Glucoregulation	-0.162

Table 3.3: Top five upregulated and downregulated pancreas genes for each exposure cluster. The top five average upregulated and downregulated pancreas-related genes for each of the five identified clusters as defined by p-value significance. Gene descriptions are obtained from the Zebrafish Information Network (ZFIN) and the National Center for Biotechnology Information (NCBI).

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
Cluster 1				
Up	*pkma*	pyruvate kinase M1/2a	involved in cellular response to insulin stimulus and glycolytic process	Type II diabetes mellitus, Glucagon signaling pathway
	pfkpa	phospho-fructokinase, platelet a	Predicted to be involved in several processes, including canonical glycolysis; fructose 1,6-bisphosphate metabolic process	Glucagon signaling pathway
	cpa1	carboxypeptidase A1 (pancreatic)	Expressed in the pancreas; pancreatic digestive enzyme.	Pancreatic secretion
	irs2b	insulin receptor substrate 2b	insulin receptor binding activity and phosphatidylinositol 3-kinase binding activity	insulin receptor binding, insulin receptor signaling pathway, positive regulation of type B pancreatic cell proliferation, Insulin signaling pathway, Type II diabetes mellitus, Insulin resistance
	ptprna	protein tyrosine phosphatase receptor type Na	involved in insulin secretion involved in cellular response to glucose stimulus and protein dephosphorylation	insulin secretion involved in cellular response to glucose stimulus, positive regulation of type B pancreatic cell proliferation, Type I diabetes mellitus
Down	*atp2a1*	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 1	Predicted to have several functions, including ATP binding activity; calcium ion binding activity; and ion transmembrane transporter activity,	Pancreatic secretion
	pfkma	phospho-fructokinase, muscle a	Human ortholog(s) of this gene implicated in glycogen storage disease VII	Glucagon signaling pathway

Continued on next page

Table 3.3 – continued from previous page

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
	atp1a1a.3	ATPase Na+/K+ transporting subunit alpha 1a, tandem duplicate 3	Predicted to have sodium:potassium-exchanging ATPase activity.	Insulin secretion, Pancreatic secretion
	ctsh	cathepsin H	Predicted to be involved in proteolysis involved in cellular protein catabolic process.	Type 1 diabetes mellitus
	dpp4	dipeptidyl-peptidase 4	Human ortholog(s) of this gene implicated in type 2 diabetes mellitus	glucagon processing
Cluster 2				
Up	*slc2a8*	solute carrier family 2 member 8	Predicted to be involved in glucose import	glucose transmembrane transport
	stat3	signal transducer and activator of transcription 3 (acute-phase response factor)	Predicted to have DNA-binding transcription activator activity.	Insulin resistance
	mlxip	MLX interacting protein	Predicted to have DNA-binding transcription factor activity.	Insulin resistance
	cacna1sa	calcium channel, voltage-dependent, L type, alpha 1S subunit, a	Predicted to have high voltage-gated calcium channel activity.	Insulin secretion
	cdx4	caudal type homeobox 4	Involved in several processes, including animal organ development.	pancreatic D cell differentiation, pancreas development
Down	*ppdpfa*	pancreatic progenitor cell differentiation and proliferation factor a	Involved in cell fate specification; cell population proliferation; and exocrine pancreas development.	exocrine pancreas development
	ctsbb	cathepsin Bb	Human ortholog(s) of this gene implicated in diabetes mellitus	Tropical calcific pancreatitis
	otop1	otopetrin 1	Exhibits cholesterol binding activity	cellular response to insulin stimulus
	dhrs9	dehydrogenase/ reductase (SDR family) member 9	Expressed in the digestive system	exocrine pancreas development

Continued on next page

Table 3.3 – continued from previous page

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
	slc27a2b	solute carrier family 27 member 2b	Predicted to have catalytic activity.	Insulin resistance
Cluster 3				
Up	*dpp4*	dipeptidyl-peptidase 4	Human ortholog(s) of this gene implicated in type 2 diabetes mellitus	glucagon processing
	tm4sf4	transmembrane 4 L six family member 4	Involved in negative regulation of pancreatic A cell differentiation; negative regulation of type B pancreatic cell development; and pancreatic epsilon cell differentiation	endocrine pancreas development, pancreatic epsilon cell differentiation, negative regulation of type B pancreatic cell development, negative regulation of pancreatic A cell differentiation
	atp1a2a	ATPase Na+/K+ transporting subunit alpha 2a	Predicted to have sodium:potassium-exchanging ATPase activity.	Insulin secretion, Pancreatic secretion
	cd36	CD36 molecule	Human ortholog(s) of this gene implicated in diabetes mellitus	Insulin resistance
	stx1a	syntaxin 1A	Predicted to have SNAP receptor activity and SNARE binding activity.	Insulin secretion
Down	*prkar1ab*	protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1) b	Predicted to have 3',5'-cyclic-GMP phosphodiesterase activity.	cellular response to glucagon stimulus, Insulin signaling pathway
	cyp26a1	cytochrome P450, family 26, subfamily A, polypeptide 1	development	pancreas development
	isl2b	ISL LIM homeobox 2b	Involved in exocrine pancreas development	exocrine pancreas development
	klf11a	Kruppel like factor 11a	Human ortholog(s) of this gene implicated in maturity-onset diabetes of the young type 7 and type 2 diabetes mellitus.	Maturity onset diabetes of the young (MODY)
	uqcc2	ubiquinol-cytochrome c reductase complex assembly factor 2	regulation of insulin secretion	regulation of insulin secretion

Continued on next page

Table 3.3 – continued from previous page

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
Cluster 4				
Up	*hspd1*	heat shock 60 protein 1	Human ortholog(s) of this gene implicated in glucose intolerance	Type I diabetes mellitus
	atf6b	activating transcription factor 6 beta	Predicted to have DNA-binding transcription factor activity and cAMP response element binding activity.	Insulin secretion
	pygl	phosphorylase, glycogen, liver	Human ortholog(s) of this gene implicated in glycogen storage disease	Insulin signaling pathway, Glucagon signaling pathway, Insulin resistance
	prkcz	protein kinase C, zeta	Involved in several processes, including animal organ development.	Insulin signaling pathway, Type II diabetes mellitus, Insulin resistance
	foxo1a	forkhead box O1 a	Is expressed in digestive system.	Insulin signaling pathway, Glucagon signaling pathway, Insulin resistance
Down	*itpr1b*	inositol 1,4,5-trisphosphate receptor, type 1b	Predicted to have inositol 1,4,5 trisphosphate binding activity and inositol 1,4,5-trisphosphate-sensitive calcium-release channel activity.	Glucagon signaling pathway, Pancreatic secretion
	nfkbiaa	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha a	Involved in negative regulation of hematopoietic progenitor cell differentiation.	Insulin resistance
	pck1	phosphoenolpyruvate carboxykinase 1 (soluble)	Human ortholog(s) of this gene implicated in type 2 diabetes mellitus	cellular response to insulin stimulus, cellular response to glucagon stimulus, Insulin signaling pathway, Glucagon signaling pathway, Insulin resistance
	ceacam1	CEA cell adhesion molecule 1	unknown	insulin catabolic process
	pck2	phosphoenolpyruvate carboxykinase 2 (mitochondrial)	Predicted to be involved in several processes, including cellular response to hormone stimulus; gluconeogenesis; and monocarboxylic acid metabolic process.	cellular response to insulin stimulus, Insulin signaling pathway, Glucagon signaling pathway, Insulin resistance
Cluster 5				
Up	*calm3a*	calmodulin 3a (phosphorylase kinase, delta)	Predicted to enable calcium ion binding activity.	Insulin signaling pathway, Glucagon signaling pathway

Continued on next page

Table 3.3 – continued from previous page

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
	pparaa	peroxisome proliferator-activated receptor alpha a	Expressed in the digestive system.	Glucagon signaling pathway, Insulin resistance
	rpl3	ribosomal protein L3	Involved in exocrine pancreas development.	exocrine pancreas development
	rps6	ribosomal protein S6	Predicted to be a structural constituent of ribosome.	Insulin signaling pathway
	celf1	cugbp, Elav-like family member 1	Organ development.	pancreas development, pancreas morphogenesis
Down	*atp1a1a.3*	ATPase Na+/K+ transporting subunit alpha 1a, tandem duplicate 3	Predicted to have sodium:potassium-exchanging ATPase activity.	Insulin secretion, Pancreatic secretion
	ccka	cholecystokinin a	Predicted to be involved in digestion.	Insulin secretion, Pancreatic secretion
	srsf6b	serine and arginine rich splicing factor 6b	Human ortholog(s) of this gene implicated in proliferative diabetic retinopathy	negative regulation of type B pancreatic cell apoptotic process
	zfand3	zinc finger, AN1-type domain 3	Involved in endocrine pancreas development.	endocrine pancreas development
	atp1b3b	ATPase Na+/K+ transporting subunit beta 3b	Predicted to enable ATPase activator activity.	Insulin secretion, Pancreatic secretion

3.3.3 Pancreas Gene Co-expression Network Analysis Identifies Key Regulatory Genes

To identify the key genes driving the distinct pancreatic responses among the chemical exposures, we constructed gene co-expression networks. This approach aimed to uncover the network of gene interactions and determine the influential genes governing the observed pathway enrichments and functional changes. Pancreas gene co-expression networks were created for all chemical exposures and for each specific cluster of chemicals. This method allowed for the identification of modules of highly correlated genes, facilitating the analysis of

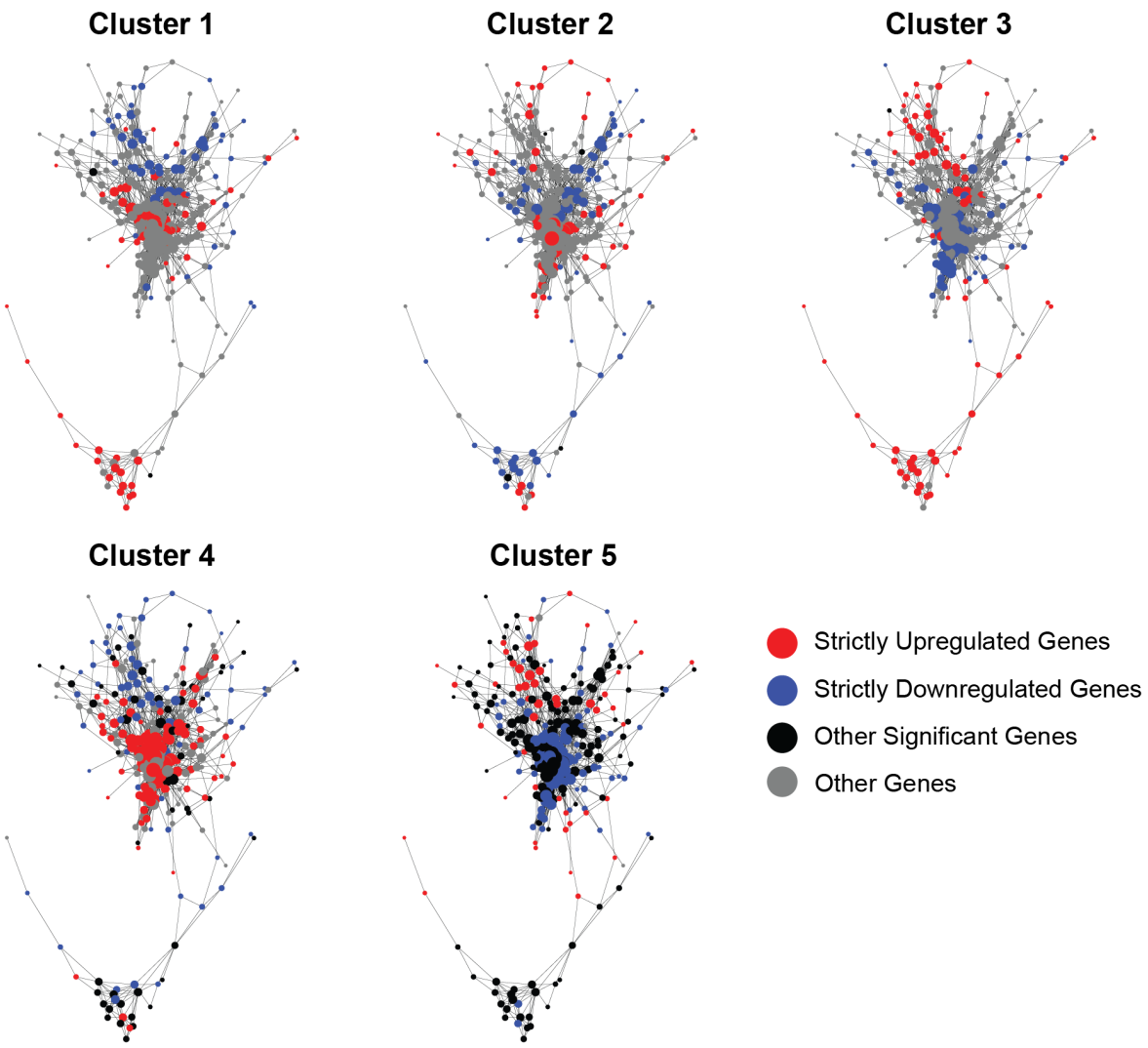

Figure 3.4: Complete pancreas gene co-expression network for 53 chemical exposures. The pancreas gene co-expression networks where nodes represent genes and edges represent the correlation between any two genes for all chemical exposures. The strictly up- and down-regulated pancreas genes by cluster are shown on the network along with other significant genes (not strictly up- or downregulated) and other non-significantly differentially expressed genes by cluster.

gene networks and the detection of central hub and bottleneck genes within these networks. A network "hub" is defined as a gene with a high node degree, indicating a high number and strength of co-expressed genes. A "bottleneck" gene is defined as a gene with a high betweenness degree that indicates a gene that connects clusters of co-expressed genes and often acts as a key regulatory gene.

A global gene co-expression network for all chemical exposures is shown in Figure 3.4. This network represents the correlation between genes across all 53 unique chemical exposures. To visualize the differences in differential gene expression between clusters, we overlayed the strictly up- and downregulated genes on the global gene co-expression network for comparison. Larger nodes indicate higher node degrees, revealing that Clusters 4 and 5 have distinct differences in up and down regulated genes governing the network behavior. Specifically, the majority of high-degree nodes in Cluster 4 are upregulated, while those in Cluster 5 are down regulated (Figure 3.4). Interestingly, Cluster 1 did not present with differential expression significance with regard to hub nodes, as indicated by the majority of up- and downregulated genes located on the outer portions of the network.

Centrality analysis on the global network resulted in the top 5 "hub" (degree) and top 5 "bottleneck" (betweenness) nodes in Table 3.4. The gene with the top degree and therefore the most highly correlated gene in the global network is *apc* (APC regulator of WNT signaling pathway), which is also reflected in the top betweenness nodes. Unsurprisingly, nearly all of the hub and bottleneck genes in the global co-expression network are related to binding and signaling. Synaptic and signaling genes (*rims2a*, *atp2b3a*, *nlgn2a*) are among the top hub nodes that highlights the importance of these genes in the regulation of pancreatic functions such as insulin release following chemical exposure [121]. The regulatory genes *cdx1b* and *gata5* are among the top bottleneck nodes, linking larger modules of co-expressed genes together. This indicates their role as key regulators in the global network, particularly in response to the larger set of all chemical exposures.

Cluster-specific gene co-expression networks are shown in Figure 3.5. For each cluster, the network represents genes (nodes) and the co-expression (edges) of any two genes for the chemical exposures within a cluster. Centrality analysis was conducted on each cluster co-expression network to identify the key genes within the networks and how they vary by cluster and is shown in Table 3.5. The strictly up- and downregulated genes are overlayed

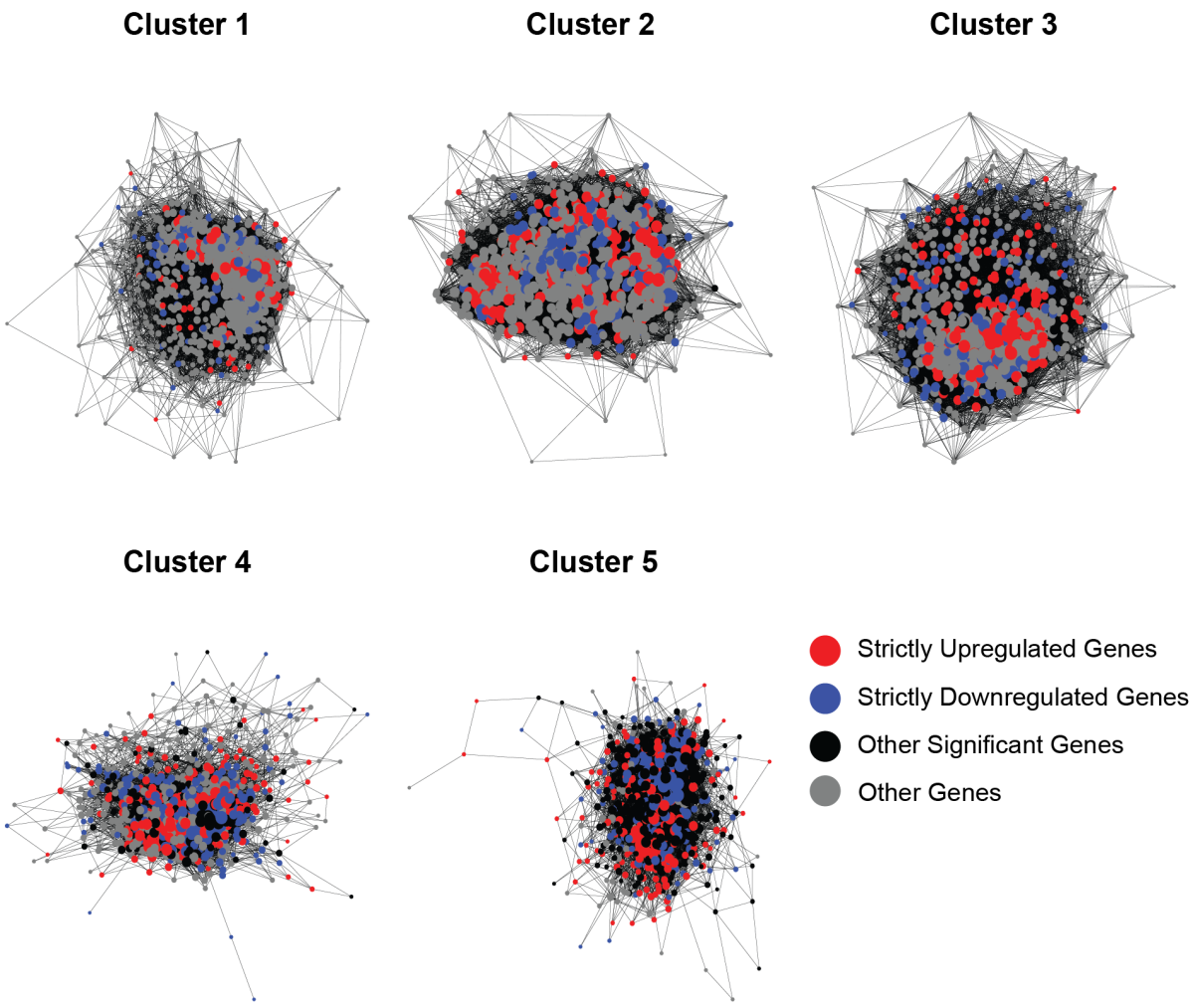

Figure 3.5: Cluster-specific pancreas gene co-expression network. The pancreas gene co-expression networks for each individual cluster. Nodes represent pancreas genes and edges represent the correlation between any two genes for all chemical exposures within the cluster. The strictly up- and downregulated genes by cluster are shown on the network along with other significant genes (not strictly up- or downregulated) and other non-significantly differentially expressed genes by cluster.

onto the co-expression networks to demonstrate the spread of significantly expressed genes for the chemical exposures within a cluster. Compared to Figure 3.4, the networks in Figure 3.5 contain more nodes and edges, indicating higher gene co-expression within clusters than across all chemical exposures (Table 3.6). More specifically, the number of nodes in the cluster-specific networks (Figure 3.5) have a 2-fold increase in nodes and a minimum of a

Table 3.4: Top degree and betweenness nodes in the global co-expression network. The top five "hub" genes (with the highest degree) and the top five "bottleneck" genes (with the highest betweenness) in the global pancreas gene co-expression network. Gene descriptions from ZFIN and NCBI are included to explain the roles of these central genes in the network.

	Gene Symbol	**Gene Name**	**Gene Description**	**Member of Pathway(s)**
Degree	apc	*APC regulator of WNT signaling pathway*	Predicted to have beta-catenin binding activity; microtubule binding activity; and protein kinase regulator activity.	pancreas development, exocrine pancreas development, endocrine pancreas development
	rims2a	*regulating synaptic membrane exocytosis 2a*	Predicted to have ion channel binding activity.	Insulin secretion
	atp2b3a	*ATPase plasma membrane Ca2+ transporting 3a*	Predicted to be involved in regulation of cytosolic calcium ion concentration.	Pancreatic secretion
	nlgn2a	*neuroligin 2a*	Predicted to have neurexin family protein binding activity and signaling receptor activity.	insulin metabolic process
	gad1b	*glutamate decarboxylase 1b*	Predicted to have glutamate decarboxylase activity and pyridoxal phosphate binding activity.	Type I diabetes mellitus
Betweenness	myo1cb	*myosin Ic, paralog b*	Predicted to have ATP binding activity.	positive regulation of cellular response to insulin stimulus
	cdx1b	*caudal type homeobox 1 b*	Exhibits DNA-binding transcription factor activity and sequence-specific DNA binding activity.	pancreas development
	gata5	*GATA binding protein 5*	Exhibits DNA binding activity and DNA-binding transcription factor activity.	pancreas development
	snap25b	*synaptosome associated protein 25b*	Involved in clustering of voltage-gated sodium channels and locomotory behavior.	Insulin secretion
	apc	*APC regulator of WNT signaling pathway*	Predicted to have beta-catenin binding activity; microtubule binding activity; and protein kinase regulator activity.	pancreas development, exocrine pancreas development, endocrine pancreas development

5-fold increase in links from the global network (Figure 3.4) across all cluster. This indicates a strong relationship between the genes expressed among the chemical exposures withing a cluster as links indicate strong co-expression between any two genes and only genes with more than one connection are included within the network.

Table 3.5: Top degree and betweenness nodes for each cluster-specific gene co-expression network. The top five "hub" genes and the top five "bottleneck" genes for each cluster-specific pancreas gene co-expression network. Gene descriptions from ZFIN and NCBI are provided to elaborate on the functions and significance of these genes within each cluster.

	Gene Symbol	**Gene Name**	**Gene Description**	**Member of Pathway(s)**
Cluster 1				
Degree	*atp1a3a*	ATPase Na+/K+ transporting subunit alpha 3a	Predicted to enable P-type sodium:potassium-exchanging transporter activity. Expressed in the pancreas primordium.	Insulin secretion, Pancreatic secretion
	pax6a	paired box 6a	Human ortholog(s) of this gene implicated in glucose intolerance.	insulin metabolic process, Maturity onset diabetes of the young
	atp1b2a	ATPase Na+/K+ transporting subunit beta 2a	Predicted to contribute to sodium:potassium-exchanging ATPase activity. Is expressed in the pancreatic bud.	Insulin secretion, Pancreatic secretion
	cacna1sb	calcium channel, voltage-dependent, L type, alpha 1S subunit, b	Predicted to have high voltage-gated calcium channel activity.	Insulin secretion
	nr5a5	nuclear receptor subfamily 5, group A, member 5	Exhibits DNA binding activity and DNA-binding transcription factor activity. Is expressed in the digestive system.	Maturity onset diabetes of the young
Betweenness	*ptpn11a*	protein tyrosine phosphatase non-receptor type 11a	Exhibits protein tyrosine phosphatase activity.	Insulin resistance
	calm2b	calmodulin 2b, (phosphorylase kinase, delta)	Predicted to enable calcium ion binding activity.	Insulin signaling pathway, Glucagon signaling pathway
	atp1b1b	ATPase Na+/K+ transporting subunit beta 1b	Predicted to enable ATPase activator activity.	Insulin secretion, Pancreatic secretion
	ascl1b	achaete-scute family bHLH transcription factor 1b	Predicted to enable DNA-binding transcription factor activity. Involves in embryonic development and endocrine pancreas development.	endocrine pancreas development
	ppp4ca	protein phosphatase 4, catalytic subunit a	Predicted to enable protein threonine phosphatase activity.	Glucagon signaling pathway
Cluster 2				

Continued on next page

Table 3.5 – continued from previous page

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
Degree	*pask*	PAS domain containing serine/threonine kinase	Predicted to be involved in intracellular signal transduction and negative regulation of glycogen biosynthetic process.	regulation of glucagon secretion
	sorl1	sortilin-related receptor, L(DLR class) A repeats containing	Predicted to be involved in post-Golgi vesicle-mediated transport.	insulin receptor recycling
	fscn1a	fascin actin-bundling protein 1a	Enables signaling receptor binding activity.	endocrine pancreas development
	creb3l4	cAMP responsive element binding protein 3-like 4	unknown	Insulin secretion, Glucagon signaling pathway, Insulin resistance
	celf1	cugbp, Elav-like family member 1	Exhibits mRNA 3'-UTR binding activity.	pancreas development, pancreas morphogenesis
Betweenness	*pfklb*	phosphofructokinase, liver b	Predicted to be involved in several processes, including canonical glycolysis and fructose 1,6-bisphosphate metabolic process.	Glucagon signaling pathway
	trib3	tribbles pseudokinase 3	Predicted to have mitogen-activated protein kinase kinase binding activity.	Insulin resistance
	pik3cd	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit delta	Predicted to have 1-phosphatidylinositol-3-kinase activity	Insulin signaling pathway, Type II diabetes mellitus, Insulin resistance
	nras	NRAS proto-oncogene, GTPase	Predicted to have GDP binding activity; GTP binding activity; and GTPase activity.	Insulin signaling pathway
	phka1a	phosphorylase kinase, alpha 1a (muscle)	Human ortholog(s) of this gene implicated in glycogen storage disease and glycogen storage disease IXd.	Insulin signaling pathway, Glucagon signaling pathway
Cluster 3				
Degree	*srebf1*	sterol regulatory element binding transcription factor 1	Involved pancreas regeneration and type B pancreatic cell development.	type B pancreatic cell development, endocrine pancreas development, response to glucagon, pancreas regeneration, Insulin signaling pathway, Insulin resistance

Continued on next page

Table 3.5 – continued from previous page

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
	flot1b	flotillin 1b	Predicted to be involved in several processes, including positive regulation of cell-cell adhesion mediated by cadherin.	Insulin signaling pathway
	sfrp1a	secreted frizzled-related protein 1a	Predicted to enable Wnt-protein binding activity.	negative regulation of canonical Wnt signaling pathway involved in controlling type B pancreatic cell proliferation
	crebbpb	CREB binding protein b	Predicted to enable chromatin DNA binding activity.	Glucagon signaling pathway
	fkbp1b	FKBP prolyl isomerase 1B	Predicted to enable peptidyl-prolyl cis-trans isomerase activity.	negative regulation of insulin secretion involved in cellular response to glucose stimulus
Betweenness	*ryr2a*	ryanodine receptor 2	Predicted to enable calcium channel activity.	type B pancreatic cell apoptotic process, Insulin secretion, Pancreatic secretion
	prkcbb	protein kinase C, beta b	Predicted to have several functions, including androgen receptor binding activity.	Insulin secretion, Insulin resistance, Pancreatic secretion
	cyc1	cytochrome c-1	Predicted to enable electron transfer activity and metal ion binding activity. Expressed in the digestive system	response to glucagon
	nphp4	nephronophthisis 4	Regulated the Wnt signaling pathway	determination of pancreatic left/right asymmetry
	nkx6.2	NK6 homeobox 2	Involved in pancreatic A cell differentiation and type B pancreatic cell differentiation.	type B pancreatic cell differentiation, pancreatic A cell differentiation, endocrine pancreas development
Cluster 4				
Degree	*pparg*	peroxisome proliferator-activated receptor gamma	Predicted to enable RNA polymerase II cis-regulatory region sequence-specific DNA binding activity and nuclear receptor activity. Expressed in the digestive system.	Type 2 diabetes mellitus

Continued on next page

Table 3.5 – continued from previous page

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
	pik3r3b	phosphoinositide-3-kinase, regulatory subunit 3b (gamma)	Predicted to be involved in insulin receptor signaling pathway.	insulin receptor signaling pathway, Insulin signaling pathway, Type II diabetes mellitus, Insulin resistance
	calml4a	calmodulin-like 4a	Predicted to have calcium ion binding activity and enzyme regulator activity.	Insulin signaling pathway, Glucagon signaling pathway
	cpa1	carboxypeptidase A1 (pancreatic)	Expressed in the pancreas; pancreatic digestive enzyme.	Pancreatic secretion
	gata5	GATA binding protein 5	Exhibits DNA binding activity and DNA-binding transcription factor activity.	pancreas development
Betweenness	*slc4a4b*	solute carrier family 4 member 4b	Predicted to have sodium:bicarbonate symporter activity.	Pancreatic secretion
	pparab	peroxisome proliferator-activated receptor alpha b	Predicted to enable DNA-binding transcription repressor activity, RNA polymerase II-specific; RNA polymerase II cis-regulatory region sequence-specific DNA binding activity; and nuclear receptor activity.	Glucagon signaling pathway, Insulin resistance
	pik3r3b	phosphoinositide-3-kinase, regulatory subunit 3b (gamma)	Predicted to be involved in insulin receptor signaling pathway.	insulin receptor signaling pathway, Insulin signaling pathway, Type II diabetes mellitus, Insulin resistance
	ogt.1	O-linked N-acetylglucosamine (GlcNAc) transferase, tandem duplicate 1	Enables protein N-acetylglucosaminyltransferase activity.	Insulin resistance
	atp2b4	ATPase plasma membrane Ca2+ transporting 4	Predicted to enable P-type calcium transporter activity and PDZ domain binding activity.	Pancreatic secretion
Cluster 5				
Degree	*nlgn2a*	neuroligin 2a	Predicted to have neurexin family protein binding activity and signaling receptor activity.	insulin metabolic process
	gad1b	glutamate decarboxylase 1b	Predicted to have glutamate decarboxylase activity and pyridoxal phosphate binding activity.	Type I diabetes mellitus

Continued on next page

Table 3.5 – continued from previous page

	Gene Symbol	Gene Name	Gene Description	Member of Pathway(s)
	prkcbb	protein kinase C, beta b	Predicted to have several functions, including androgen receptor binding activity.	Insulin secretion, Insulin resistance, Pancreatic secretion
	atp2b3b	ATPase plasma membrane Ca2+ transporting 3b	Predicted to be involved in regulation of cytosolic calcium ion concentration.	Pancreatic secretion
	tcf7l2	transcription factor 7 like 2	Involved in several processes, including exocrine pancreas development. Used to study type 2 diabetes mellitus.	exocrine pancreas development, negative regulation of type B pancreatic cell apoptotic process, Type 2 diabetes mellitus
Betweenness	*ildr2*	immunoglobulin-like domain containing receptor 2	Involved in endoderm development and pancreas development.	pancreas development
	pla2g12a	phospholipase A2, group XIIA	Predicted to enable phospholipase A2 activity.	Pancreatic secretion
	fscn1a	fascin actin-bundling protein 1a	Enables signaling receptor binding activity.	endocrine pancreas development
	capn10	calpain 10	Predicted to act upstream of or within cellular response to insulin stimulus.	cellular response to insulin stimulus, type B pancreatic cell apoptotic process, positive regulation of type B pancreatic cell apoptotic process, Type 2 diabetes mellitus
	gata6	GATA binding protein 6	Enables DNA binding activity.	Pancreatic agenesis and congenital heart disease; Pancreatic hypoplasia diabetes heart disease; Yorifuji-Okuno syndrome

Cluster 1, which has been previously shown to be characterized by dysregulation of insulin secretion, contains the hub genes *atp1a3a* and *atp1b2a* and bottleneck gene *atp1b1b* which are all involved in ATPase activator activity. Production and maintenance of ATP is crucial for the proper functioning of pancreatic beta cells, as ATP is essential for the energy-dependent processes involved in insulin synthesis and secretion [122]. High co-expression of these ATP-related genes within the network further emphasize the impact paraben chemical exposures in Cluster 1 have on insulin dysregulation.

Table 3.6: Number of nodes and links in the network models. The number of total nodes, edges, and significant nodes that correspond to significantly differentially expressed genes are given for the global pancreas gene expression network (Figure 3.3) and the cluster specific pancreas gene expression network (Figure 3.4).

	Nodes	**Links**	**Strictly Upregulated Nodes**	**Strictly Downregulated Nodes**	**Other Significant Nodes**	**Other Non-Significant Nodes**
Global Pancreas Gene Co-Expression Network						
Cluster 1	280	1459	60	53	2	165
Cluster 2	280	1459	56	66	4	154
Cluster 3	280	1459	76	61	1	142
Cluster 4	280	1459	90	63	52	75
Cluster 5	280	1459	48	81	147	4
Cluster Specific Pancreas Gene Co-Expression Network						
Cluster 1	687	27,142	89	92	2	504
Cluster 2	687	29,123	151	120	8	408
Cluster 3	687	33,964	161	122	10	394
Cluster 4	666	7,329	156	185	105	220
Cluster 5	668	10,449	160	167	296	45

Dysregulation of glucagon signaling and insulin sensitivity characterized the top up- and down- regulated pathways and genes in Cluster 2 (Tables 3.2 and 3.3). Here, within the Cluster 2 gene co-expression network, several hub and "bottleneck" genes further explain this gene expression behavior, especially because the majority of pesticide/herbicide chemicals were within Cluster 2. The top hub genes (*pask*, *sorl1*, and *fscn1a*) highlight the importance of intracellular signaling and transport in maintaining insulin sensitivity. Meanwhile, bottleneck genes such as (*pfklb*, *trib3*, and *pik3cd*) are crucial regulators of key metabolic pathways including glycolysis, further emphasizing their roles in glucose homeostasis.

The altered pathways and genes in Cluster 3 indicated impaired pancreatic development (Tables 3.2 and 3.3). The co-expression network reveals the genes most highly co-expressed with remaining genes that play critical roles in altered developmental processes. Among the top hub genes related to development of the pancreas are *srebf1*, *flot1b*, and *sfrp1a*. The gene *rebf1* encodes a protein that is directly involved in pancreatic cell development and regeneration, while *flot1b* and *sfrp1a* are important for cell adhesion and Wnt signaling,

respectively, all crucial for pancreatic tissue integrity and development. The bottleneck genes *nphp4* and *nkx6.2* regulate pathways necessary for pancreatic cell differentiation and development such as the Wnt signaling pathway, highlighting their importance in forming functional pancreatic tissues.

Cluster 4 primarily contains chemical exposures from HOCs and PFAS. Interestingly, top hub and bottleneck genes in the Cluster 4 co-expression network include *pparg* and *pparab*, respectively. Both genes are a member of the Peroxisome Proliferator-Activated Receptors (PPARs) which are transcription factors that play crucial roles in lipid and glucose metabolism and are often studied using zebrafish to investigate obesity and other metabolic diseases [123]. Our previous studies have demonstrated that PPAR agonists and antagonists can exert significant effects on gene expression, embryonic morphology, and pancreas development, indicating the importance of PPAR alterations to chemical exposures in Cluster 4 [124, 65]. Furthermore, bottleneck genes *ogt.1* and *atp2b4* are involved in post-translational modification and calcium ion transport, respectively, processes vital for pancreatic cell function and insulin secretion.

Glucoregulation was among the top pathway alterations observed in Cluster 5. Hub genes within the co-expression network include *gad1b*, *prkcbb*, and *tcf712* which directly relate to insulin secretion and glucose metabolism, indicating genes in Cluster 5 are most commonly associate with genes involved in glucoregulation. Additionally, among the top bottleneck genes are *atp2b3b* and *capn10*, both involved in calcium signaling which is essential for the cellular response to insulin.

Table 3.7: Mean cross-validation (CV) metrics and standard deviation for machine learning models. The average cross-validation accuracy, precision, recall, and F1-score with associated standard deviations for different machine learning models used to predict cluster membership and shared pancreatic toxicity profile based on chemical properties for a 4-fold cross-validation.

Model	***Accuracy***		***Precision***		***Recall***		***F1-score***	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Random Forest	**0.74**	0.08	**0.79**	0.10	**0.74**	0.08	**0.73**	0.09
XGBoost	0.70	0.04	0.76	0.04	0.70	0.04	0.69	0.05
Support Vector Machine	0.68	0.12	0.69	0.15	0.68	0.12	0.66	0.14
Multiclass Logistic Regression	0.68	0.10	0.73	0.10	0.68	0.10	0.68	0.10

3.3.4 Machine Learning Predicts Chemical Property Impacts on Pancreatic Function

To gain further insight into the relationship between chemical properties and their impact on pancreatic function, we employed supervised machine learning. This approach allowed us to develop a model that could predict the cluster membership (and consequently the functional impact) of a chemical exposure based solely on its chemical properties. The model feature space includes chemical properties obtained from the EPA CompTox dashboard and experimentally relevant properties such as micromolar concentration, zebrafish age at RNA collection, exposure duration, repeated exposure (static vs. repeated dosing), and fish strain.

Machine learning model selection and performance was conducted to determine the best performing model for the given dataset. Using a stratified 4-fold cross validation (CV), the following machine learning models were tested for comparison: random forest, XGBoost, SVM, and multiclass logistic regression. Average model accuracy CV performance using a complete feature space are illustrated in Table 7 with random forest outperforming other models with an accuracy of 74% and F1-score of 73%. We also performed hyperparameter tuning to optimize the random forest model's performance. The ability of the random forest method to predict cluster membership using chemical properties without gene expression data as members of the feature space indicates a relationship between chemical properties

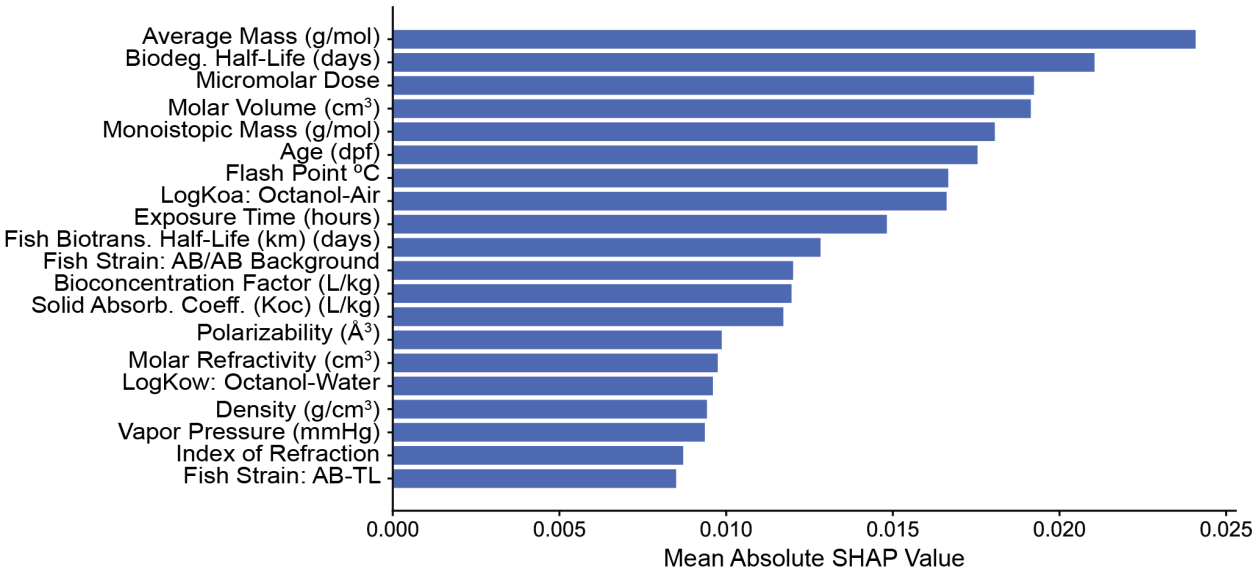

Figure 3.6: SHAP values for feature importance in predicting cluster membership. Mean absolute SHAP values for each chemical property feature used in the random forest model. Higher SHAP values indicate greater importance in predicting the cluster membership of chemical exposures. The most influential features include average mass, biodegradation half-life, and micromolar dose.

and cluster outcomes.

To investigate the chemical and biological properties that had the greatest impact on cluster membership, feature importance was conducted on the random forest model by measuring the SHAP values for each feature across each validation fold. The mean absolute SHAP values are reflected in Figure 3.6 with larger absolute SHAP values indicating higher feature importance. While average mass, biodegradation half-life, and micromolar dose ranks as the most influential features, the contribution of the top 15 features was relatively similar, with SHAP values ranging from just below 0.010 to a maximum of 0.025. This suggests that multiple chemical and biological properties contribute to cluster differentiation (Figure 3.6).

These influential features likely contribute to the observed functional differences between clusters. For example, the shorter biodegradation half-life in Cluster 1 suggests faster degradation, potentially resulting in a lesser impact on pancreatic function compared to other clusters. This is further supported by the few up- and downregulated genes observed in

the pancreas gene co-expression network (Figure 3.5). The average mass among chemical exposures, the top feature influencing cluster outcomes, varies widely among all clusters, indicating significant mass differences across cluster memberships. LogKoa, the octanol-air partition coefficient, also plays a key role in cluster assignment, whereas water solubility, which was not identified as a top feature, remains relatively consistent across clusters (Figure 3.7).

3.4 Discussion

The zebrafish model has been extensively used by the scientific community to investigate the effects of environmental pollutants, pharmaceuticals, and other chemicals on embryonic development. The goal of this study was to aggregate publicly available data on zebrafish embryos to identify chemicals that elicit similar altered pancreatic pathway responses and to investigate the chemical properties influencing these outcomes. We developed an analytical pipeline to integrate transcriptomic responses from zebrafish exposed to 53 distinct chemicals from various sources, minimizing experimental and technical variations for downstream analyses. Our orthology-based methodology allowed for the assessment of differential gene expression and pathway responses relevant to pancreatic development and function, even for pathways not currently annotated in zebrafish. By employing clustering and co-expression analyses, we identified groups of chemical exposures that distinctly impact the pancreas. Furthermore, we used machine learning to determine the chemical properties most influential in the distinct pancreatic behaviors observed in the clustering results. Our findings provide significant insights into the molecular mechanisms by which different chemical exposures alter pancreatic pathways and gene expression profiles, with machine learning revealing chemical properties that may drive these observed outcomes.

Our clustering analysis revealed distinct groups of chemical exposures that induce specific

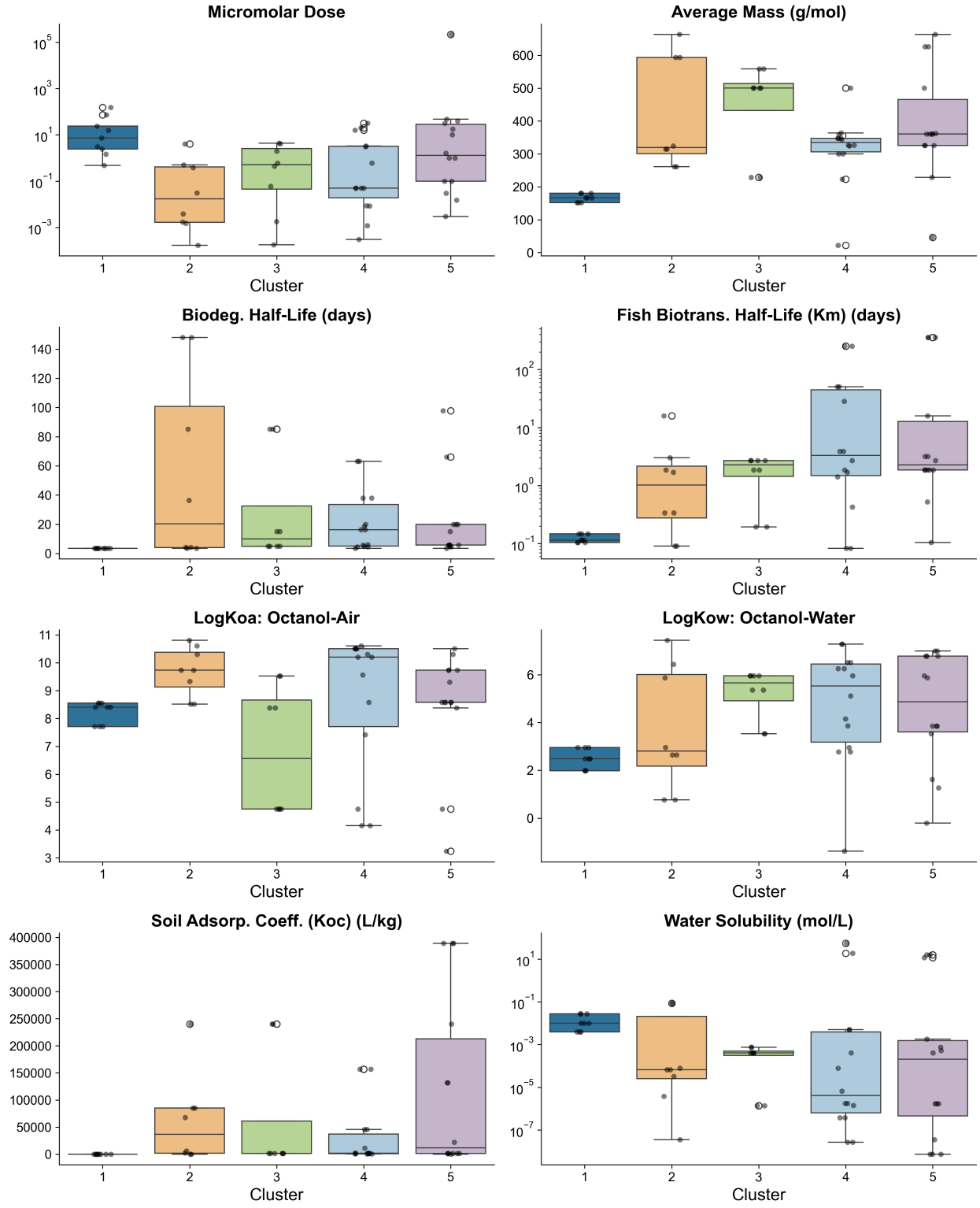

Figure 3.7: Box plots of key chemical features by cluster. Box plots illustrating the distribution of key chemical features across the different clusters.

pancreatic pathway alterations, showing that chemical class alone does not dictate these changes. Instead, clusters reflected groups of exposures resulting in similar pancreatic processes, such as insulin dysregulation, signaling, development, disease, and endocrine disruption. Parabens, forming Cluster 1, uniquely impacted insulin regulation pathways, aligning with studies linking paraben exposure to diabetes in humans [125]. Cluster 2, consisting mostly of pesticide/herbicide exposures and several pharmaceuticals, showed a common disruption of glucagon signaling, insulin sensitivity, and gene expression of *slc2a8* (*Glut8*), a known glucose transporter in humans and zebrafish. Notably, CBD and THC in this cluster affect PPAR signaling, which is directly related to glucoregulation and insulin sensitivity [126, 84]. Cluster 3 included PFOS and BPA exposures, both known endocrine disruptors, with significant alterations in pancreatic development pathways. Previous studies have shown PFOS hinders pancreatic development [65, 67] and BPA affects pancreas development in other models [127]. Clusters 4 and 5, containing a mix of chemical types, highlighted broader impacts on pancreatic development and glucoregulation. Cluster 4 was characterized by disruptions in pathways related to exocrine pancreatic insufficiency and insulin secretion although the magnitude of log odds ratio and up/down regulation was small as indicated by the colors in Figure 3.2. Cluster 5 exhibited significant alterations in cellular responses to insulin and glucagon stimuli, indicating extensive impacts on pancreatic function and glucose metabolism.

The global pancreas gene co-expression network analysis further elucidated the regulatory genes driving the differential responses following chemical exposure. Key hub and bottleneck genes for the global co-expression network, such as *apc* and *cdx1b*, were identified, highlighting their roles in binding and signaling processes crucial for maintaining pancreatic processes and function in response to chemical exposure. Interestingly, *apc* has been shown to be a top bottleneck gene in a zebrafish chemical exposure network for a different set of chemicals including Flame Retardant Chemicals and AHR2 activators [18]. Beyond *apc*'s involvement as a key regulator of the WNT signaling pathway, it is a known tumor suppressor impacting

oncogenesis, suggesting its sensitivity as a marker for toxicological effects. Additionally, mutations of *apc* are implicated in the development of pancreatic cancer [128]. The recurrent identification of *apc* as a key regulator in our study and the aforementioned study highlights the observed relationship between chemical exposures, *apc*, and the potential implications of *apc* alterations including pancreatic cancer.

The chemical clustering outcomes are further supported by the cluster-specific pancreas gene co-expression networks. Each cluster network obtained an increase in nodes and connections when compared to the global pancreas gene co-expression network, which correspond to a high number of co-expressed genes (Table 3.6). The key driver genes (Table 3.5) of the tightly connected networks (Figure 3.5), complement the top upregulated and downregulated genes and pathways identified through the hierarchical clustering analysis. Interestingly, Cluster 4 contained PPARs (*pparg* and *pparab*) as key driver genes in the cluster-specific pancreas gene co-expression networks while Cluster 5 contained *ppara* as a top upregulated gene. It has previously been shown that PPAR signaling can disrupt pancreatic development in the zebrafish embryo, underscoring the potential for altered pancreatic processes following exposure to the chemicals in these clusters [124]. Additionally, Clusters 4 and 5 primarily consist of known endocrine-disrupting chemicals (EDCs), including halogenated organic compounds. These chemicals are recognized for their ability to act as PPAR agonists, influencing PPAR signaling pathways [129]. The identification of PPARs as central nodes in the cluster-specific network reinforces the importance of these receptors in mediating the effects of chemical exposures on the pancreas.

Computational methods and frameworks are continually emerging to investigate critical genes contributing to a disease state or conditions of interest [130, 73, 131]. In the computational toxicology context, understanding the chemical properties that explain the gene associations observed can provide critical information to build upon for toxicological evaluation [132]. The analytical framework developed for this study is the first of its kind to

integrate a meta-analysis of zebrafish transcriptomic datasets following chemical exposure and chemical properties that impact the transcriptomic observations using machine learning. The random forest model, achieving an accuracy of 74%, identified key chemical properties such as average mass, and biodegradation half-life, and micromolar dose as significant predictors of cluster membership. SHAP analysis further revealed that multiple chemical and biological properties contribute to cluster outcomes, rather than a single dominant feature driving classification. By quantifying the relative importance of these features, this method helps identify key chemical properties influencing pancreatic pathway alterations while maintaining model interpretability. This integrated approach not only uncovers key chemical features influencing the pancreas but also sets a foundation for future toxicity prediction models as the dataset scales.

This approach represents a crucial avenue for scrutinizing whole-organism datasets to assess target organ toxicity, particularly in light of the evolving paradigm towards New Approach Methodologies (NAMs) [133]. As the field progresses, there's a growing emphasis on incorporating more high-throughput organisms like zebrafish, C. elegans, and Drosophila into toxicity screens to assess bioactivity *in vivo* [134, 135, 136]. Leveraging these organisms allows for comprehensive evaluations of chemical effects on entire biological systems, capturing nuanced responses that may be missed in traditional *in vitro* assays [13]. With the expanding repertoire of whole-organism datasets, researchers can gain deeper insights into the complex interplay between chemical exposures and biological responses, even at the organ level as discussed here in the pancreas context. A key future direction is to expand this research to other organ systems and incorporate chemical mixtures, allowing for a more comprehensive understanding of how diverse tissues respond to complex chemical exposures. This data-driven analysis of untargeted experimental datasets facilitates more informed decision-making in toxicological assessments and advances the development of safer chemicals and pharmaceuticals.

The modeling pipeline and approach should be considered in light of some limitations. Firstly, this study was dependent on probing publicly available data currently listed in GEO, which confines the chemicals able to be included within the study. Additionally, there are a myriad of experimental approaches that have been developed to study organ-specific transcriptomic responses, such as single-cell RNA-sequencing and spatial transcriptomics. These advanced techniques offer high-resolution insights into cellular and spatial transcriptomic landscapes; however, they come with significant barriers. The cost and accessibility of these techniques are substantial, requiring specialized equipment, laboratory expertise, extensive computational resources, and time as they are inherently low-throughput. In contrast, our approach leverages existing datasets and applies computational techniques to overcome these challenges and assess the impact of chemical exposures on a target organ. The ability to screen a large set of chemicals efficiently is a significant advantage, facilitating comprehensive toxicological evaluations and advancing our understanding of chemical-induced alterations in pancreatic processes.

In conclusion, our study demonstrates the power of integrating transcriptomic data from a variety of zebrafish chemical exposures with advanced analytical techniques to uncover the molecular underpinnings of chemical-induced pancreatic alterations. The identification of specific chemical exposure clusters and the associated pancreas gene expression changes provide valuable insights into the pathways affected by different types of chemicals. Moreover, the use of machine learning to highlight key chemical properties influencing these pathways underscores the potential for predictive models in toxicology. As we continue to expand and refine this dataset with the growing body of research, the methodologies and findings presented here will serve as a robust foundation for future research, ultimately contributing to the larger toxicological evaluation efforts.

Chapter 4

Associations Between Birth Defect Incidence and Maternal, Sociodemographic, and Environmental Factors in California from 2018-2019: A Computational Approach

The motivation for Chapter 4 stems from a desire to extend our previously developed methodologies from Chapter 3 to a population-level dataset in order to better understand how environmental conditions relate to observed birth defects across California. The original data use agreement for the California Department of Public Health (CDPH) was established as part of a master's thesis project by Arielle Beaulieu of the Sant Laboratory, which investigated the associations between drinking water contaminants and birth defect outcomes using

traditional statistical analyses [137]. Although that initial study did not establish a clear link, likely due to the rarity of birth defects and the resulting small sample sizes in an associative study design, it highlighted the need for a more comprehensive dataset and advanced analytical approaches. Motivated by these limitations, this chapter expands the scope of environmental exposures beyond drinking water contaminants to include factors such as air quality and pesticide exposure, and incorporates additional pregnancy and labor-delivery complications. By enhancing the data sources and refining the modeling strategy, we aim to overcome the challenges posed by the low frequency of birth defect outcomes. Recognizing that birth defects are inherently multifactorial, we hypothesized that an integrative approach combining multilayer network analysis with machine learning would be fruitful for elucidating the complex relationships between environmental exposures, demographic factors, and birth defect incidence. The results of this Chapter have been submitted for publication with Arielle Beaulieu as a co-author for her contribution in establishing the initial data use agreement and organization of the CDPH dataset.

Birth defects are a leading cause of infant morbidity and mortality, with a significant portion of cases lacking a clear etiology. Understanding the complex interactions between environmental, biological, and demographic factors is critical for predicting birth defect risks and informing prevention strategies. In this study, we leverage a multilayer network modeling approach combined with machine learning to uncover these interactions and identify key predictors of birth defects. We analyzed data from a cohort of 552,688 mothers in California, incorporating 101 outcome variables from four distinct domains: characteristics (parental and infant), physical and social environment, pregnancy complications, and labor and delivery complications. A multilayer network model was employed to assess the co-occurrence of each variable both within and across these four domains and to systematically identify the influential variables that should be utilized in the construction and refinement of the machine learning models, to enhance predictive accuracy. Machine learning techniques, including Random Forest and XGBoost, were applied to identify key predictors of birth defect

risk and quantify their relative contributions, providing insights into the most influential factors driving these outcomes. The multilayer network model revealed significant interactions between various health, environmental, and demographic factors, with mothers who had birth defects showing increased co-occurrence of variables across layers, supporting the compounded effects and multifactorial nature of birth defect outcomes. Machine learning models exhibited strong predictive performance for birth defect outcomes (AUROC = 0.88), based on a reduced feature space of influential variables identified through multilayer network analysis, improving both interpretability and efficiency. Among the top model predictors, impaired water bodies emerged as a key environmental factor. This study demonstrates the value of combining multilayer network modeling with machine learning to explore complex relationships between birth defect risk factors. The approach provides a more holistic understanding of birth defect etiology, with implications for targeted interventions and policy development. By identifying key environmental and health-related predictors, this work contributes to the growing body of research on birth defect prevention and risk assessment.

4.1 Introduction

Although the rates of infant mortality, including those attributable to birth defects, have steadily declined since the 1970s in the United States, birth defects remain the leading cause of infant mortality, accounting for 19.5% of cases [138]. Despite these improvements, disparities persist, with certain populations experiencing higher rates of birth defects and infant mortality. Studies have highlighted how factors such as race and ethnicity can influence these outcomes, with some minority groups showing an increase in infant mortality rates in recent years [138, 139]. While approximately 30% of all birth defects have a known cause, such as genetics, the remaining 65-75% of all birth defects have an unknown etiology and may be attributable to factors such as infections and the environment [140].

Many epidemiological studies have established associations between birth defects and socioeconomic and environmental factors, suggesting these play a critical role in their development [19, 141]. For example, environmental pollutants such as drinking water contaminants and air pollution have been linked to birth defects [142, 3, 4]. However, the existing literature often presents varied findings, reflecting the complexity of studying multifactorial problems. The sheer number of variables, spanning health, environmental, and socioeconomic domains, complicates efforts to disentangle their contributions to adverse outcomes.

Traditional studies on birth defect outcomes typically focus on single factors or isolated domains, such as socioeconomic or environmental factors, limiting their ability to capture the complex interplay of these variables. One common limitation is data availability, which constrains researchers' ability to investigate broader relationships comprehensively, especially as it relates to rare birth defect outcomes [143]. However, California's extensive datasets on birth defect outcomes, combined with information on regional socioeconomic and environmental factors, provide a unique opportunity to explore these interactions. Prior studies have used aggregated measures to examine these relationships, such as indices combining socioeconomic and environmental burdens, to examine these relationships [19]. While helpful, such methods may oversimplify the intricate interconnections among variables.

To address these challenges and capture the complexity of multifactorial data, advanced modeling techniques are needed. Network modeling is a mathematical modeling strategy utilized to link and investigate related variables and entities using graph theory [144]. These models have been applied successfully in various biological fields, such as identifying highly correlated gene expression patterns, protein interactions, or developmental anomalies in a model organism [145, 146, 75, 147]. Multilayer networks extend this approach by incorporating multiple domains of data [148, 149]. This allows for the analysis of both within-domain and cross-domain relationships, making it particularly well-suited for studying birth defect outcomes, which arise from a combination of factors including parental health conditions,

socioeconomic indicators, and environmental exposure. Although multilayer network models are powerful for understanding variable interactions, its combination with a predictive model to understand the true influence of a variable on model outcomes can yield impactful insights.

Machine learning is a powerful tool for predictive modeling, capable of uncovering nonlinear interactions between factors that might otherwise go undetected. This approach has been applied to predict birth defect outcomes and other adverse health events using diverse datasets, including health records and environmental exposure data [150, 151, 152]. Machine learning models also provide insights into which variables most strongly influence prediction and have been shown to offer interpretable outputs for various health concerns and public health initiatives [153, 154]. Machine learning model predictions are often sensitive to the feature space and model development as highly correlated features can decrease interpretability. The information from the multilayer network modeling strategy can be utilized to inform a more robust machine learning model that utilizes knowledge of variable interactions.

In this study, we integrate multilayer network modeling and machine learning to investigate the multifactorial contributors to birth defects. Using a population-based dataset of live birth outcomes and a dataset capturing California's physical and social environment, we hypothesized that incorporating a broad set of domains, including labor and delivery complications, pregnancy complications, parent giving birth and infant characteristics including demographic factors, and environmental exposures, would improve prediction accuracy. Moreover, we aimed to identify key predictors that have the largest impact on birth defect outcomes, with particular attention to the interactive and multidimensional nature of these variables. Our findings demonstrate the utility of combining multilayer networks and machine learning to better understand the complex relationships underlying birth defects. By highlighting key predictors and their interactions, this work provides actionable insights to help reduce risk and inform public health strategies aimed at preventing birth defects.

4.2 Materials and Methods

4.2.1 Data

Data related to pregnancy outcomes and the parent giving birth were obtained from the California Department of Public Health California Comprehensive Master Birth File (CCMBF) through a data use agreement. CCMBF maintains information of all live births occurring in California based on information entered on birth certificates as well as out of state births from California residents. Data related to social and environmental factors across California was obtained from the California Environmental Protection Agency California Communities Environmental Health Screening Tool (CalEnviroScreen) program which is publicly available on the CalEnviroScreen webpage [155]. All outcomes in the CCMBF related to labor and delivery complications and pregnancy complications were included in the curated dataset. Specific demographic information was selected for the parent giving birth and the newborn based on curation of the authors from the literature. Identification of a parent giving birth's physical and social environment was determined based on the census tract listed within the CCMBF file and was removed upon matching with the CalEnviroScreen dataset to ensure no personally identifiable information is reported.

4.2.2 Population and Data Preparation

Data used for this study was taken from years 2018-2019 to ensure optimal data completeness across CCMBF and CalEnviroScreen as CalEnviroScreen does not have exhaustive longitudinal data. While CalEnviroScreen 3.0 provides environmental factor data from 2011–2019, birth survey data underwent significant improvements in 2018, resulting in more complete and comprehensive information from that point onward. There were 906,545 birth records during this timeframe that have aggregated outcomes on abnormal newborn conditions, labor

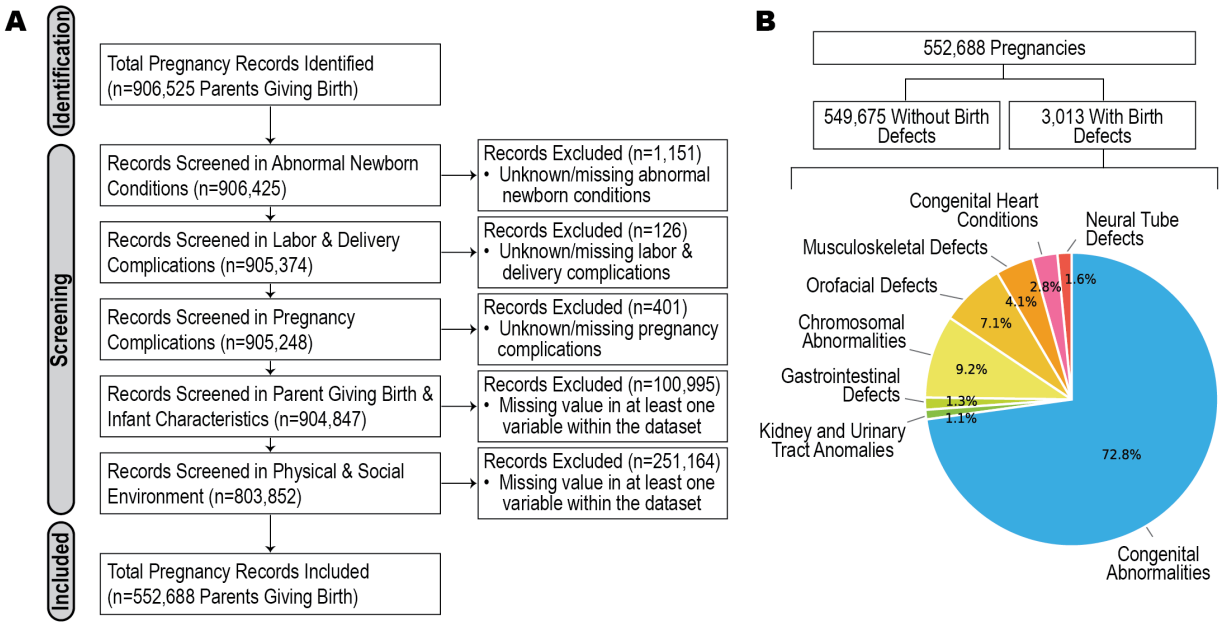

Figure 4.1: PRISMA flow diagram summarizing study inclusion criteria from the California Department of Public Health Birth Monitoring Program and the California Environmental Protection Agency California Communities Environmental Health Screening Tool (CalEnviroScreen). (A) PRISMA chart illustrating the study inclusion criteria and the removal of observations based on screening criteria and (B) the categories of observed defects within the birth defect cohort.

and delivery complications, pregnancy complications, parent giving birth and infant characteristics, and the physical and social environment (Figure 4.1A). For each aggregated set, records containing missing data was removed from the final dataset during screening (Figure 4.1A). The majority of removed records were from missing data within the CalEnviroScreen physical and social environment due to incomplete information across the chosen timeframe. The final dataset comprised of 552,688 birth records with 3,013 birth defects recorded in the abnormal newborn condition's subsection of the CCMBF file for having one or more of the following conditions: congenital abnormalities, kidney and urinary tract anomalies, gastrointestinal defects, chromosomal abnormalities, orofacial defects, musculoskeletal defects, congenital heart conditions, and neural tube defects (Figure 4.1B). Greater than 70% of the study population with a recorded defect was identified to have a congenital abnormality (Figure 4.1B). The variables included in this study that are further tested for predictive power

are categorized into one of four variable subsets: labor and delivery complications (25 variables), pregnancy complications (24 variables), parent giving birth and infant characteristics (29 variables), and physical and social environment (26 variables).

4.2.3 Multilayer Network Model

Co-occurrence Identification

The co-occurrence of any two variables was computed using statistical tests tailored to the type of variable pair, ensuring that statistical assumptions were met. Each variable represents a health, demographic, or environmental outcome coming from one of our four variable subsets: labor and delivery complications, pregnancy complications, parent giving birth and infant characteristics, and physical and social environment, with a comprehensive list provided in Table C.1. Continuous data were log-transformed to satisfy normality requirements where applicable. For comparisons between categorical variables, a chi-squared test was used to measure association. For continuous variables, Pearson's correlation was computed. For comparisons involving one continuous and one categorical variable, the Kruskal-Wallis test was applied. Statistical significance was defined as $p < 0.01$, and to ensure observed effects were meaningful, an additional effect size threshold was implemented. Effect sizes specific to each statistical test were used to confirm that statistically significant results also reached at least a "small" effect threshold. The effect size measures employed were Cramér's V for the chi-squared test, Pearson's r for correlation, and eta-squared (η^2) for the Kruskal-Wallis test. As raw effect size outcomes are not necessarily comparable across effect size type, we take known intervals for effect size threshold unique to the specific measure to determine size [156]. Thresholds for small, medium, and large effect sizes are provided in Table 4.1. Two variables were considered to co-occur if they met both criteria: $p < 0.01$ and an effect size of at least "small."

	Cramer's V	Pearson	Eta-Squared
Small	0.05	0.3	0.02
Medium	0.06	0.5	0.04
Large	0.14	0.7	0.09

Table 4.1: Effect size minimum thresholds for identification as a small, medium or large effect depending on statistical test (chi-squared, Pearson correlation, and Kruskal-Wallis. Effect sizes are utilized to ensure a statistically significant result ($p < 0.01$) maintains sufficient effect.

Network Properties

Multilayer network models represent mathematical graphs that can be utilized to represent and investigate the co-occurrence of variables for the parents giving birth that have a newborn with a birth defect and those without. A multilayer network model incorporating the co-occurrence of each variable pair for each cohort (without-defects and with-defects) was represented using a unique multilayer network model that was built on foundational mathematical properties and the statistical significance of co-occurrence, as defined in the proceeding section. To ensure comparability across effect size measurements, the true effect size was binned into small (s), medium (m), and large (l) effect given values of 0.33, 0.66, and 1 respectively. The development of the multilayer network model for both the without-defects and with-defect cohort is described below with visualizations of network behavior conducted using Python NetworkX [102].

A single-layer network model can be represented mathematically by a graph $G = (N,\ E,\ \omega)$, where N represents the set of nodes, $E \subset N \times N$ represents the set of edges, and $\omega : N \times N$ is the edge weight mapping function such that each edge $e_{ij} \in E$ has a weight ω_{ij} [103]. For each variable subset, a single layer network model was constructed where each node $n_{ij} \in N$ represents a variable, each edge $E \in N \times N$ with associated weight w_{ij} represents the co-occurrence between any two variables within the variable subset, measured by the effect size ($s = 0.33$, $m = 0.66$, $l = 1$). The edges of the network are un-directed. The mathematical graph of this single layer network model can be represented by an adjacency matrix A of

size $N \times N$ such that $A_{ij} = \omega_{ij}$ if there is an edge between nodes n_i and n_j and $A_{ij} = 0$ otherwise. The multilayer network extension of the single-layer is a four-layer network, where each layer corresponds to one of the four variable subsets [157]. The multilayer network is represented mathematically by a multilayer graph $\mathcal{G}$ consisting of the 4 single-layer networks G_k with associated adjacency matrices A_k for $k = 1, \ldots, 4$. Each layer is fully connected to the set of nodes in the corresponding layers. Therefore, we will obtain inter-layer adjacency matrices $A_{l,m}$ for each pair of graphs G_l and G_m for $l, m = 1, \ldots, 4,\ l \neq m$ that specifies the edges between nodes in different layers. Each edge between layer nodes has an associated weight computed from the co-occurrence effect size. The structure of the final multilayer network model $\mathcal{G}$ can then be described by a supra-adjacency matrix $\mathcal{A}$ containing all node and weighted edge information for the network in tensor form.

Density of the Multilayer Network Model

To quantify the connectivity of the multilayer network models, we calculated the weighted edge density for each layer and the overall network. Edge density is a measure of the proportion of observed connections relative to the total possible connections, providing insights into the network's structural characteristics and variable interactions [103]. By incorporating edge weights into this calculation, we accounted for the strength of co-occurrence relationships between variables.

The weighted edge density $D_{l,m}$ for any intra- or inter-layer is computed using Eq. 4.1.

$$D_{l,m} = \frac{\sum A_{l,m}}{\zeta}, \quad \zeta = \begin{cases} \frac{1}{2} N_l \left(N_l - 1\right) \text{ for } l = m \text{ (intra} - \text{layer)} \\ N_l N_m \text{ for } l \neq m \text{ (inter} - \text{layer)} \end{cases} \tag{4.1}$$

where $A_{l,m}$ is the adjacency matrix the adjacency matrix containing edge weights and variable co-occurrence values for intra-layers ($l = m$) and inter-layers ($l \neq m$) N_l and N_m represent the

number of nodes (variables) within the respective layers. The numerator $\sum A_{l,m}$ aggregates the weights of all edges in the respective layer or between layers, while ζ represents the total number of possible connections.

The computed edge densities provided a measure of how densely connected each layer and the overall network were, weighted by the strength of variable co-occurrence. This calculation offered insights into the degree to which variables co-occur within and between variable subsets [158]. Comparisons of edge densities between the without-defects and with-defects cohort networks highlighted differences in connectivity patterns that may be associated with birth defect outcomes.

Clustering of the Multilayer Network Model

To uncover meaningful patterns and relationships within the variables that make up the multilayer network model, we applied the Louvain clustering algorithm, a widely used method for detecting community structures in networks. Louvain clustering is an optimization-based algorithm that maximizes modularity, a measure of the density of connections within clusters compared to connections between clusters [159]. This method is particularly well-suited for multilayer network models, as it can accommodate the complex interplay between intra-layer and inter-layer edges.

The clustering analysis was conducted in Python 3.9 using the NetworkX package, applied independently to each cohort-specific multilayer network (without-defects and with-defects) to identify distinct communities of co-occurring variables. Before clustering, the multilayer network's supra-adjacency matrix was transformed into a modularity-maximizing graph representation, which is crucial for ensuring that the detected communities reflect natural groupings based on the strength of relationships between variables [159]. The Louvain algorithm iteratively grouped variables into communities by optimizing modularity until convergence,

identifying the optimal number of clusters with maximal modularity. The resulting clusters represent groups of variables with strong co-occurrence relationships within and across layers, capturing potential shared mechanisms or interactions. These clusters offer valuable insights into the factors and relationships associated with birth defect outcomes.

Identification of Top Variables in the Multilayer Network Model

Centrality analysis is a method for uncovering the topology of a multilayer network model, identifying the nodes corresponding to variables within each subset that are most influential and frequently co-occur with other variables in the model. This analysis highlights the variables that play pivotal roles in the network's structure and may have significant biological or environmental implications. To evaluate the importance of the nodes in the network, we utilized three centrality metrics: eigenvector centrality, degree centrality, and betweenness centrality. These complementary measures provide a comprehensive assessment of node influence by capturing different aspects of connectivity and structural importance.

Eigenvector centrality measures a node's influence based on the importance of its neighbors, assigning higher scores to nodes connected to other high-scoring nodes [103]. This metric is particularly useful for identifying variables that are part of tightly connected and influential clusters within the network. The eigenvector centrality of node $n_i \in N$ in the multilayer network model is measured using the eigenvalues of the supra-adjacency matrix as $C_E(i) = \frac{1}{\lambda}\sum_{j\in N}\omega_{i,j}x_j$ where $\omega_{i,j}$ is the strength of co-occurrence between nodes i and j, x_j is the centrality of node j, and λ is the maximum positive eigenvalue of $\mathcal{A}$. Nodes with high eigenvector centrality are considered central to the overall network dynamics.

Degree centrality measures the number of direct connections a node has, reflecting its immediate level of activity or prominence within the network. For weighted networks, degree centrality is calculated as the sum of the weights of all edges connected to the node as

$C_D(i) = \sum_{j \in N} \omega_{i,j}$ [160]. Nodes with high degree centrality are key hubs within the network, known to co-occur with many other variables either within or cross layers.

Betweenness centrality identifies nodes that serve as critical bridges or intermediaries within the network. It quantifies the number of shortest paths between pairs of nodes that pass through a given node [103]. The betweenness centrality of a node i is defined as $C_B(i) = \sum_{s,t \in V} \frac{\sigma(s,t|i)}{\sigma(s,t)}$ where $\sigma(s,t)$ is the total number of shortest paths between nodes s and t, and $\sigma(s,t|i)$ is the number of those paths that pass through node i. Nodes with high betweenness centrality are crucial for maintaining connectivity and facilitating interactions between different regions of the network.

Centrality measures were computed for each node in the multilayer network model, considering both intra- and inter-layer edges. By ranking nodes based on these centrality metrics, we identified key variables within each variable subset and across the entire network that drive co-occurrence patterns. For this analysis, top nodes were selected based on eigenvector centrality with the other centrality measures also reported, as this metric prioritizes variables that co-occur strongly with other influential variables, capturing the tightly connected structures relevant to the biological and environmental contexts of the network. Comparisons of centrality rankings between the without-defects and with-defects cohort networks revealed differences in the top-ranked variables, potentially shedding light on mechanisms or interactions associated with birth defect outcomes.

4.2.4 Machine Learning Model

Feature Preparation

As described in the Methods - Data and Methods - Population and Data Preparation sections, data for each birth record and parent giving birth, spanning four diverse variable sets, were

aggregated for analysis. Variables with more than two categories were one-hot encoded to represent binary outcomes (yes/no) for each category. Continuous variables were scaled to have a mean of 0 and a standard deviation of 1 to reduce the impact of outliers and improve model comparisons. To investigate the importance of the feature space on model predictions, three feature sets were prepared:

1. *All Variables*: Includes all variables within the dataset.
2. *Without Environment*: Excludes variables in the physical and social environment subset.
3. *Top Variables*: Features derived from the top 30 variables identified from the with-defects cohort based on eigenvector centrality.

Modeling

Supervised machine learning models were developed to predict birth defect outcomes based on features representing labor and delivery complications, pregnancy complications, parent giving birth and infant characteristics, and physical and social environment factors. Interpretability, handling of imbalanced data, and sensitivity to feature structure were prioritized. The following models were evaluated: logistic regression, random forest, and XGBoost [161].

The data was split into an 80/20 train/test set for training and evaluation. Hyperparameter optimization was conducted during training using 10-fold cross-validation, with the area under the receiver operating characteristic curve (AUROC) as the optimization metric. Grid search was used to select the best hyperparameters for each model (Table C.2). Model performance was evaluated using the following metrics:

- Area Under the Receiver Operating Curve (AUROC): Measures the model's ability

to distinguish between positive (birth defect) and a negative (no defect) classes by plotting the true positive rate versus the false positive rate at different thresholds.

- Area Under the Precision Recall Curve (AUPR): Measures the model's ability to identify positive instances while minimizing false positives by plotting precision versus recall at different thresholds.

- Precision: Proportion of correctly identified positive cases among all positive predictions. Higher precision indicates fewer false positives.

- Recall: Proportion of actual positive cases correctly identified. Higher recall indicates fewer false negatives.

- F1: Harmonic mean of precision and recall, balancing the two metrics when both are important.

The final model was evaluated on the test set using a stratified approach to account for potential class imbalance. Performance was assessed across the five metrics for each of the three feature sets, providing insights into the impact of feature selection on model predictions.

Model Explanations

Understanding the contributions of individual features to model predictions is critical for interpreting the relationship between input variables and birth defect outcomes. Tree based methods such as those utilized in this study are known to capture complex relationships that make the identification of feature importance increasingly difficult. To explain the complex interactions and importance, we utilized SHAP (SHapley Additive exPlanations) values to quantify the importance of each feature in the machine learning model while accounting for nonlinear relationships.

SHAP values explain a model's prediction by attributing the contribution of each feature to a predicted outcome both at the local (individual) level and can be aggregated to determine a global measure of feature importance (average absolute SHAP value) [162]. These values are calculated while considering the marginal contribution of each feature across all possible subsets of features, giving a fair measure of importance that may increase of decrease the likelihood of a positive outcome (e.g. the presence of a birth defect). We utilized SHAP values to investigate feature importance in the final trained modes for the *All Variables* and the *Top Variables* feature sets along with the importance of variable subsets on the overall model predictions. As SHAP values are inherently additive, variable subset contributions were computed by taking the sum of all SHAP values for a feature within a variable subset and normalizing to the number of total features within the variable subset.

To enhance the interpretability of SHAP values, each value was converted to a risk ratio. The conversation to a risk ratio was performed by scaling the SHAP values relative to the base prediction probability of the model following the method described in [150]. The log-odds of birth defect probability for a given sample are expressed as:

$$LO = \phi_0 + \sum_{i=1}^{d} \phi_i$$

Here, ϕ_0 represents the base SHAP value (the logit of the population prevalence, P_0), and ϕ_i represents the SHAP values attributed to features $i \in 1, \ldots d$. The predicted probability for an outcome based on the contribution of individual features is computed using the sigmoid function, the inverse of the logit:

$$P_i = S\left(\phi_0 + \phi_i\right) = \frac{1}{1 + e^{-(\phi_0 + \phi_i)}}$$

The relative risk (RR) for a single feature i is then defined as the ratio of the predicted

probability associated with the feature to the base probability:

$$RR_i = \frac{P_i}{P_0} = \frac{S(\phi_0 + \phi_i)}{S(\phi_0)}$$

To create dependence plots that demonstrate the feature value as a function of the RR across all observations, the calculated RRs were grouped into bins based on the feature values. For each bin, the mean and standard deviation of RRs were computed and plotted against the mean feature value represented as uncertainty bands. Presenting SHAP values in terms of the RR provides a more actionable interpretation of how changes in a feature's value influence the risk of birth defects while maintaining the integrity of SHAP dependence plots.

4.3 Results

4.3.1 Descriptive Statistics: Overview of Population Characteristics

To investigate the variables that may be related to an increased risk of birth defects, we aggregated data from the California Department of Public Health (CDPH) California Comprehensive Master Birth File (CCMBF) and the California Environmental Protection Agency California Environmental Screen (CalEnviroScreen) dataset. The CCMBF dataset contains detailed information about the parent giving birth-newborn pairs including demographic information and specific outcomes related to the pregnancy and abnormal newborn conditions. CalEnviroScreen contains general information about the physical and social environment by census tract in California that is matched to the parent giving birth based on the where the parent giving birth is known to reside. A detailed overview of the data preprocessing strategy and subsequent analyses is given in Figure 4.2. The final dataset used for analysis

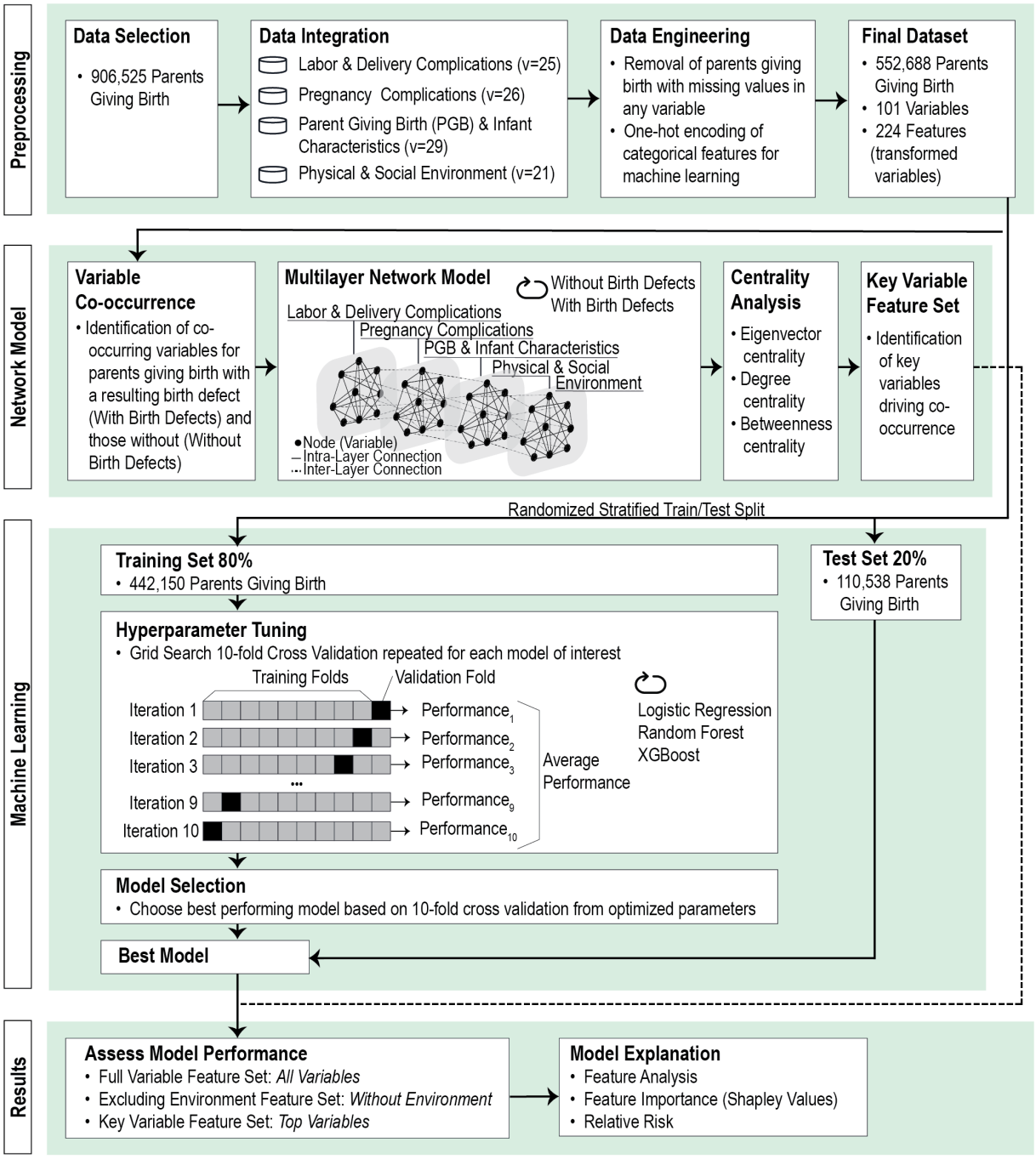

Figure 4.2: Analysis workflow for investigation of the multifaceted nature of birth defect outcomes. A concise overview of the modeling strategy utilized for the study, including data preprocessing, network modeling and machine learning implementation. A detailed description of study inclusion criteria is provided in the Methods.

after preprocessing (see Methods) is composed of 552,688 parents giving birth from 2018-2019 in the state of California with a known newborn birth defect outcome (yes/no) and 4

variable subsets: labor and delivery complications (25 variables), pregnancy complications (26 variables), parent giving birth and infant characteristics (29 variables), and physical and social environment (21 variables). A condensed representation of characteristics, sociodemographic, and other demographic information for the parent giving birth and the newborn is shown in Table 4.2 with a comprehensive depiction of all included variables in Table C.1.

Of the 552,688 parents giving birth included within the dataset, 3,013 had a newborn with an identified birth defect (less than 1%). The number of male infants in both the without birth defects and with birth defects categories were slightly larger than female infants. The mean age of the parent giving birth in both the without and with birth defects cohort remains close to 30 years of age. Recognizing the multifaceted nature of birth defect outcomes, we employed a multilayer network and machine learning approach, as illustrated in Figure 1, to explore the complex variable relationships between variables in Table S1.

Table 4.2: Characteristics [mean ± SD or n (%)] of the without-birth defects and with-birth defects cohorts comprised of the parent giving birth and newborn pairs for live births occurring in the state of California from 2018-2019. Full characteristics are presented in Table C.1 with a condensed version shown here.

	Without Birth Defects (N=549,675)	**With Birth Defects (N=3,013)**
Parent Giving Birth Characteristics		
Age	30.11 ± 5.85	29.01 ± 6.34
Weight Pre-pregnancy	155.08 ± 38.94	161.12 ± 38.16
Weight at Delivery	183.12 ± 38.54	186.51 ± 38.94
Height	63.67 ± 2.84	63.44 ± 2.81
Pre-pregnancy BMI	26.88 ± 6.40	28.12 ± 6.27
Race		
White	388837 (70.74)	2648 (87.89)
Black	32889 (5.98)	106 (3.52)
Asian	90305 (16.43)	155 (5.14)
Native Hawaiian or Other Pacific Islander	2797 (0.51)	7 (0.23)
American Indian or Alaskan Native	3232 (0.59)	16 (0.23)
Other Specified	31615 (5.75)	81 (2.69)
Multi-race		
Non-Hispanic	271626 (49.42)	745 (24.73)
Hispanic Single Race	257472 (46.84)	2201 (73.05)
Two or More Races - Any Hispanic Status	20577 (3.74)	67 (2.22)
Education		

Continued on next page

Table 4.2 – continued from previous page

	Without Birth Defects (N=549,675)	With Birth Defects (N=3,013)
≤High School/GED	204728 (37.25)	1617 (53.67)
Some College/Associate	151226 (27.51)	809 (26.85)
≥Bachelor's Degree	193721 (35.24)	587 (19.48)
Women, Infants, and Children (WIC) Program	223120 (40.59)	1587 (52.67)
Payment for Prenatal Care		
Private Insurance Company	283232 (51.53)	931 (30.90)
Medical/Other Government Support	238473 (43.38)	1902 (63.13)
Self-Pay	18336 (3.34)	111 (3.68)
Other	6744 (1.23)	6 (0.20)
No Prenatal Care	2890 (0.53)	63 (2.09)
Payment for Delivery		
Private Insurance Company	282980 (51.48)	919 (30.50)
Medical/Other Government Support	241504 (43.94)	1941 (64.42)
Self-Pay	18274 (3.32)	144 (4.78)
Other	6578 (1.20)	6 (0.20)
Medically Unattended Delivery	339 (0.06)	3 (0.10)
Final Route of Delivery		
Vaginal	380869 (69.29)	1902 (63.13)
Cesarean	168806 (30.71)	1111 (36.87)
Children Born Alive		
1-3 Children	485400 (88.31)	2579 (85.60)
≥4 Children	64275 (11.69)	434 (14.40)
Infant Characteristics		
Child Sex		
Female	268311 (48.81)	1374 (45.60)
Male	281361 (51.19)	1638 (54.36)
Birthweight (grams)	3290.89 ± 557.23	3142.44 ± 689.12
Gestation (weeks)	38.6 ± 1.91	37.96 ± 2.55
Gestation (days)	269.88 ± 35.06	270.35 ± 29.63
Plurality		
Single	533188 (97)	2911 (96.61)
Twin	16119 (2.93)	100 (3.32)
Fetal Presentation		
Cephalic	525076 (95.52)	2816 (93.46)
Breech	20343 (3.70)	171 (5.68)
Other	4256 (0.77)	26 (0.86)
APGAR Min5		
Score ≤5	4386 (0.80)	115 (3.82)
Score >5	545289 (99.20)	2898 (96.18)
APGAR Min10		
Score ≤5	1202 (0.22)	72 (2.39)
Score >5	548473 (99.78)	2941 (97.61)
Hearing Test		
Pass (Both Ears)	248064 (45.13)	346 (11.48)
Fail (One or Both Ears)	5899 (1.07)	48 (1.59)
Pending	186978 (34.02)	649 (21.54)

Continued on next page

Table 4.2 – continued from previous page

	Without Birth Defects (N=549,675)	**With Birth Defects (N=3,013)**
Test Not Available	102443 (18.64)	1860 (61.73)
Not Screened due to Medical Condition	5751 (1.05)	101 (3.35)
Refused	540 (0.10)	9 (0.30)

4.3.2 Multilayer Network Modeling: Capturing Complex Variable Interactions

To capture the complexity of variable interactions, we constructed a multilayer network model that integrates multiple categories of variables across different layers. Each layer within the network model represented a distinct variable subset, with intra-layer (within-subset) and inter-layer (between-subset) links denoting statistically significant associations between variables ($p < 0.01$ and effect size $>$ minimum threshold, Methods – Table 4.1). Nodes in the network corresponded to individual variables (e.g., age), and edges represented co-occurrence relationships. Unlike traditional statistical approaches, this multilayer network approach allows simultaneous analysis of within-subset and cross-subset interactions, providing a more comprehensive view of complex associations across distinct categories of variables in the with and without birth defects cohorts. Details on the construction of the multilayer network, including statistical thresholds and criteria for defining links are provided in Methods – Multilayer Network.

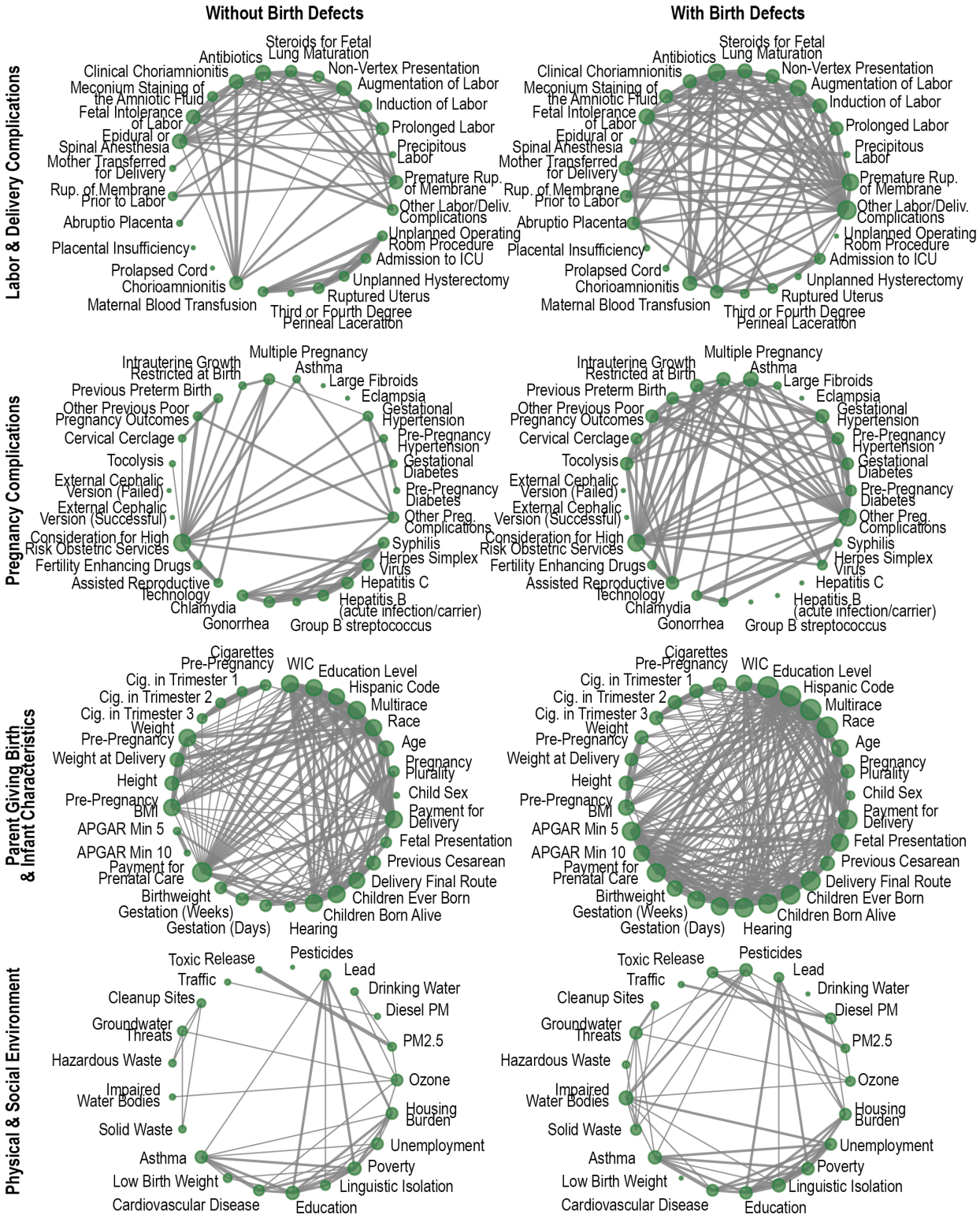

Figure 4.3: Intra-layer networks for the without- and with-birth defects cohorts across the four network layers: Labor & Delivery Complications, Pregnancy Complications, Parent Giving Birth & Infant Characteristics, and Physical & Social Environment. Larger nodes indicate a larger node degree with thicker edge links indicating strong correlations between any two nodes/variables.

Increased Cross-Domain Co-occurrence in Birth Defects Cohort Highlight Multifactorial Risk

Pregnancies resulting in birth defects displayed markedly different patterns of variable co-occurrence compared to those without, exhibiting significantly higher network connectivity and complexity. Specifically, intra-layer edge densities, which quantify the proportion of possible connections within a layer, increased significantly in the labor and delivery complications, pregnancy complications, and parent giving birth and infant characteristic subsets ($p < 0.05$, Figure 4.3 and Figure 4.4). This increase indicates higher levels of co-morbidities and complex interdependencies among these variables. In contrast, the variable connectivity within the without birth defects cohort often reflected expected population characteristics. For example, in the pregnancy complications layer, sexually transmitted diseases were strongly connected and relatively disconnected from the rest of the network, suggesting a more isolated presentation of this risk factor. However, within the birth defects cohort, the same variables exhibited greater interconnection, highlighting a disruption of typical co-occurrence patterns.

Inter-layer connections also showed a marked increase in edge density for all subset combinations in the birth defects cohort (Figure 4.4). For example, variables from the pregnancy complications subset exhibited stronger associations with variables from the physical and social environment subset, highlighting the interconnected nature of factors contributing to birth defects. In contrast, the non-birth defects cohort exhibited lower intra- and inter-layer edge densities, reflecting fewer and less complex interactions. These findings suggest that pregnancies resulting in birth defects are characterized by more intricate risk profiles involving interactions across multiple domains.

Interestingly, while intra-layer connectivity within the physical and social environment subset did not significantly increase, its inter-layer connections with other subsets (e.g., pregnancy

complications) were significantly higher in the birth defects cohort. This finding emphasizes the importance of cross-domain interactions, particularly those involving environmental factors, in understanding birth defect risks.

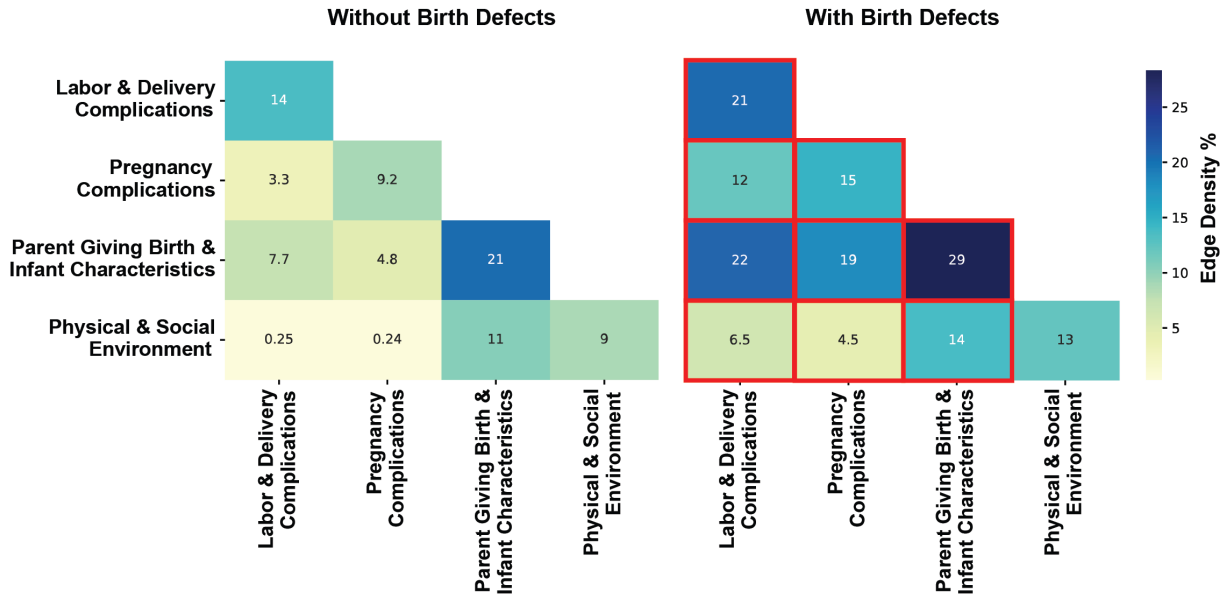

Figure 4.4: Edge density of intra- and inter- layers of the multilayer network models for the without birth defects and with birth defects cohorts. Red squares signify a statistically significant difference in edge density between the without- and with- birth defects groups ($p < 0.05$).

Distinct Clustering Patterns Reveal Key Differences in Variable Interactions Between Cohorts

Community detection within the multilayer network revealed distinct clustering patterns for variables in the birth defects compared to the non-birth defects cohorts (Figure 4.5), with detailed information on the clustering methodology provided in the Methods section. In the non-birth defects cohort, six clusters were identified, largely confined to specific variable subsets, reflecting relatively isolated relationships. Conversely, the birth defects cohort exhibited only four clusters, with extensive cross-subset interactions, indicating a more interconnected network and co-occurring variables.

Notably, several variables in the non-birth defects network were isolated due to a lack of connections (e.g., *Placental Insufficiency*, *Prolapsed Cord*, *Eclampsia*), whereas the birth defects network included only one isolated node (*Unplanned Operating Room Procedure*). Clusters in the birth defects network often spanned multiple subsets, such as Cluster 3 containing variables from physical and social environments, parent giving birth and infant characteristics, and labor and delivery complications. This integration underscores the interplay of diverse factors associated with birth defect risks.

In contrast, clusters in the non-birth defects cohort were more compartmentalized, with variables largely confined to their respective subsets. For example, Cluster 0 primarily consisting of labor and deliver complications highlights the less interconnected nature of these variables to other variable subsets in the absence of birth defects. These differences underscore the heightened complexity of risk profiles in the birth defects cohort.

Centrality Analysis Identifies Influential Variables Driving Birth Defect Risks

Centrality measures were applied to identify influential variables in the multilayer network that may play pivotal roles in pregnancy outcomes (Table 4.3). Eigenvector centrality highlighted variables with extensive connections to other well-connected nodes, degree centrality identified variables with high direct connectivity, and betweenness centrality revealed hub variables bridging different clusters.

In both cohorts, variables such as *Payment for Prenatal Care*, *Payment for Delivery*, and *Race* ranked highly in centrality measures, reflecting their pervasive influence. However, differences emerged in the specific central roles of these variables between cohorts. For example, in the birth defects network, variables such as *Antibiotics*, *APGAR Min5*, and *Hearing* were prominent, highlighting the critical role of medical interventions and testing in this cohort.

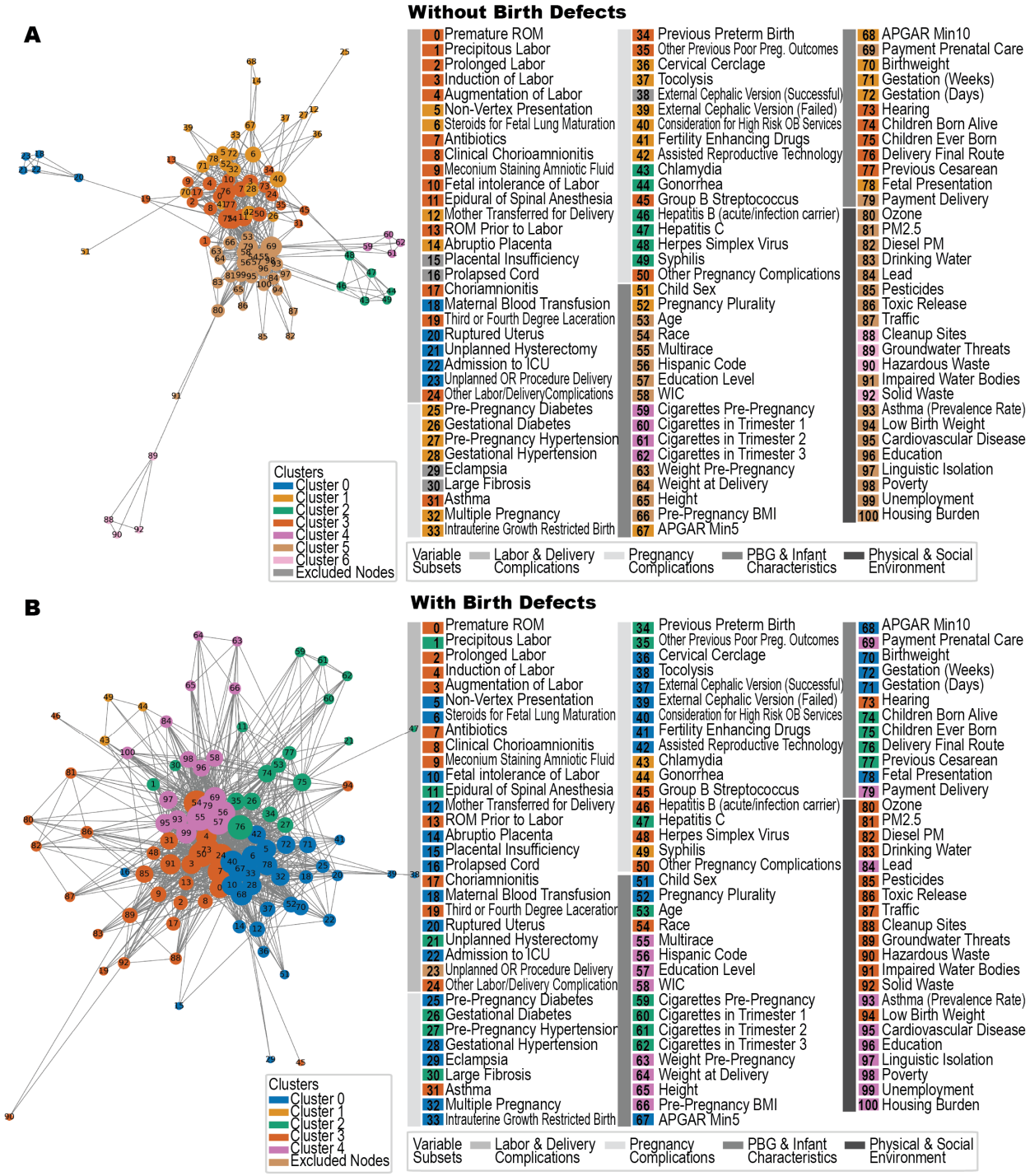

Figure 4.5: Multilayer network representation of variable co-occurrence for all variable subsets in the without-birth defects (A) and with-birth defects (B) cohorts. Variables are colored based on cluster groupings of variables indicating groups of co-occurring variables within each cohort. Excluded nodes are variables that had no connections or indications of co-occurrence with remaining variables. Abbreviations: obstetrics (OB), rupture of membrane (ROB).

The centrality analyses revealed contrasting clustering patterns between the two cohorts. In the birth defects cohort, influential variables were distributed across multiple clusters, reflecting a network with highly interconnected nodes and the complex interplay of medical, social, and environmental factors potentially contributing to birth defects. Conversely, in the without-birth defects cohort, central variables were more localized, primarily confined to Cluster 5. This cluster appeared to capture broader population-level trends, such as demographic and socioeconomic factors, rather than variables directly tied to specific medical outcomes.

These findings suggest that while pregnancies involving birth defects exhibit a multifaceted and interconnected network of risk factors, healthy pregnancies are characterized by more compartmentalized relationships, possibly reflecting a more uniform population structure driven by demographic and lifestyle factors.

4.3.3 Machine Learning: Predicting Birth Defects

We leveraged the variable subsets and knowledge of variable co-occurrence from the multi-layer network model to train machine learning models aimed at predicting parental social, medical, and environmental characteristics that may increase the incidence of birth defects. The dataset was split into 80% for training and 20% for testing, maintaining the proportion of true birth defect outcomes to address potential class imbalances. Categorical variables were one-hot encoded, converting each category into a binary feature. For example, the variable *Race* was transformed into 21 binary features, each representing a specific racial category using one-hot encoding. All transformed variables are presented in (Table C.1). Continuous variables were standardized to have a mean of 0 and a standard deviation of 1 to ensure consistency across features. Additional information on the machine learning model optimization and development is provided in Methods – Machine Learning Model.

Table 4.3: Top 20 nodes with the highest eigenvector centrality in the without and with birth defects multilayer network models. Each centrality measure is reported along with the cluster identification.

Node	Eigenvector Centrality	Degree Centrality	Betweenness Centrality	Cluster ID
Without Birth Defects				
Payment for Prenatal Care	0.257	0.442	0.139	5
Payment for Delivery	0.242	0.368	0.035	5
Multi-race	0.233	0.389	0.092	5
Number of Children Born Alive	0.221	0.358	0.059	3
Number of Children Ever Born	0.221	0.358	0.059	3
Race	0.221	0.389	0.121	5
Women, Infants and Children Program (WIC)	0.220	0.295	0.019	5
Education Level	0.216	0.316	0.038	5
Hispanic Code	0.210	0.284	0.011	5
Epidural or Spinal Anesthesia	0.198	0.305	0.035	3
Delivery Final Route	0.174	0.389	0.095	3
Education (regional)	0.163	0.189	0.002	5
Poverty	0.163	0.189	0.002	5
Asthma (prevalence rate)	0.149	0.168	0.001	5
Unemployment	0.144	0.158	0.002	5
Cardiovascular Disease	0.138	0.147	0.001	5
Pre-pregnancy BMI	0.134	0.158	0.003	5
Weight Pre-pregnancy	0.131	0.147	0.002	5
Age	0.124	0.137	0.002	5
Housing Burden	0.122	0.137	0.000	5
With Birth Defects				
Hearing	0.200	0.697	0.069	3
Other Pregnancy Complications	0.193	0.636	0.039	3
Antibiotics	0.190	0.616	0.052	3
Multi-race	0.189	0.697	0.082	4
Other Labor/Delivery Complications	0.189	0.596	0.030	3
APGAR Min5	0.184	0.616	0.075	0
Consideration for High Risk Obstetric Services	0.181	0.535	0.015	0
Education Level	0.181	0.616	0.055	4
Hispanic Code	0.178	0.586	0.036	4
Delivery Final Route	0.178	0.606	0.042	2
Race	0.169	0.646	0.080	3
Payment for Delivery	0.163	0.505	0.024	4
Steroids for Fetal Lung Maturation	0.157	0.455	0.011	0
Unemployment	0.151	0.394	0.003	4
Payment for Prenatal Care	0.149	0.465	0.019	4
Impaired Water Bodies	0.148	0.394	0.006	3
Fetal Intolerance of Labor	0.146	0.404	0.008	0
Augmentation of Labor	0.143	0.394	0.006	3
Asthma (prevalence rate)	0.142	0.354	0.002	4
Cardiovascular Disease	0.139	0.343	0.002	4

Model Selection and Performance Comparison

Figure 4.6 illustrates the five evaluation metrics (area under the receiver operating curve (AUROC), area under the precision recall curve (AUPR), precision, recall, and F1-score) for logistic regression, random forest, and XGBoost models, both with default and optimized hyperparameters, across 10-fold cross-validation using the complete feature set. Details of the hyperparameter optimization process are provided in the Methods with parameter values given in Table C.2.

This evaluation yielded ten models for each method, which were internally validated to reduce overfitting and account for variability during hyperparameter tuning. Both XGBoost and random forest models with optimized hyperparameters achieved an AUROC of 0.89. However, XGBoost consistently outperformed random forest in all other metrics, most notably in precision, which dropped significantly from 0.67 (XGBoost) to 0.40 (random forest). Logistic regression demonstrated high recall (0.97) but failed in precision, achieving near-zero values, underscoring its limitation in this context.

XGBoost demonstrated the most balanced performance across metrics, solidifying it as the top-performing model. Notably, the improvement in XGBoost's performance from default to optimized parameters was modest, with a 2% increase in AUROC. Given its robustness across all metrics, we selected XGBoost with optimized hyperparameters for subsequent analyses.

Feature Set Analysis: Physical & Social Environment Variables Increase AUROC by 3% and AUPR by 20%

To explore the impact of feature selection on predictive performance, we evaluated XGBoost models trained on three feature sets: (1) all variables, (2) all variables excluding environ-

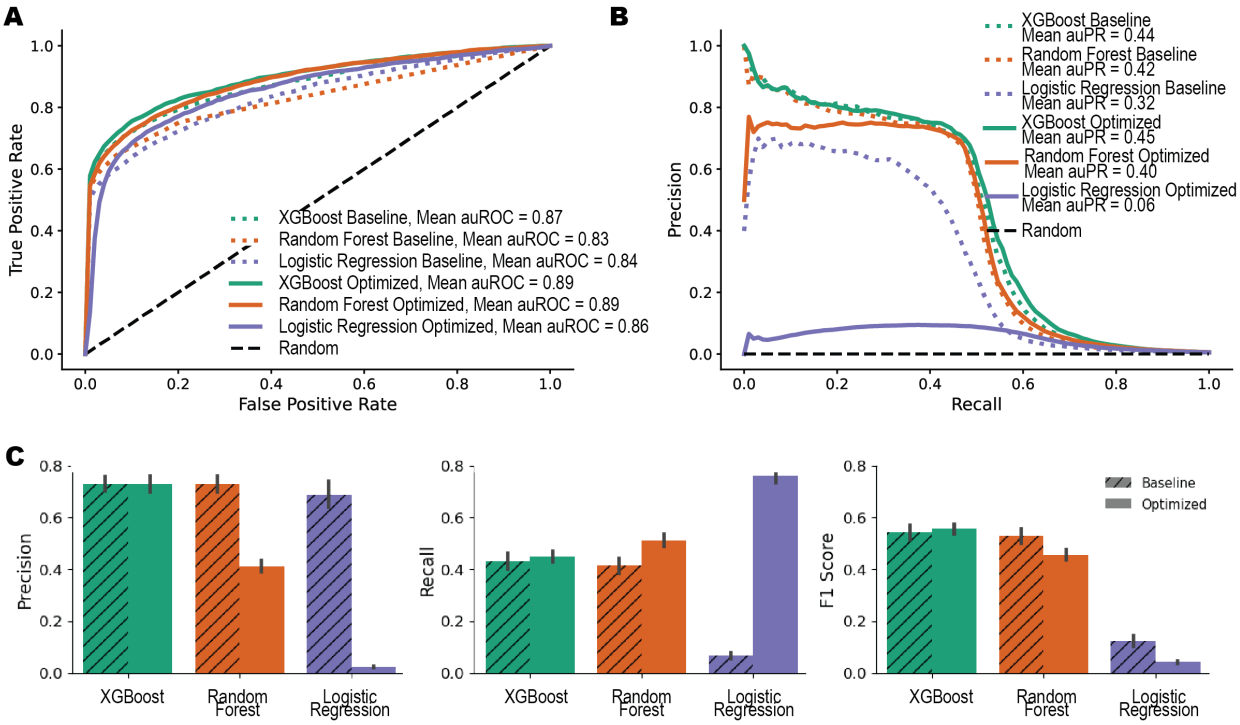

Figure 4.6: The area under the receiver operating curve (A), area under the precision-recall curve (B), and the precision, recall, and F1 score metrics (C) for the three machine learning models of interest, XGBoost, random forest, and logistic regression, evaluated on the *All Variables* feature space using 10-fold cross validation. In (C), the bars represent the standard deviation across folds for precision, recall, and F1. Baseline models (represented with dashed/hashed lines) were generated using default parameters, while optimized models were fine-tuned using hyperparameter optimization to maximize AUROC during training.

mental features, and (3) the top 30 variables identified by the multilayer network model referred to as *All Variables*, *Without Environment*, and *Top Variables* respectively (Figure 4.7). Model performance was assessed on the test set using 10-fold stratification to account for variability. The models trained on the *All Variables*, *Without Environment*, and the *Top Variable* feature sets achieved AUROCs of 0.89, 0.86, and 0.88, respectively.

Excluding environmental features drastically reduced AUPR from 0.43 to 0.23, indicating the critical role of environmental factors in improving predictive power. Interestingly, the *Top Variables* feature set maintained performance comparable to the *All Variables* feature set while significantly reducing the complexity of the feature space. By selecting only 30 variables out of the original 101, the feature space, including one-hot encoded variables, was

reduced by over 50%, decreasing computational resource requirements and highlighting the ability of the multilayer network model to effectively prioritize relevant variables.

These findings underscore the importance of including physical and social environmental factors in predictive models for birth defects. Furthermore, they demonstrate that leveraging the multilayer network for feature selection retains predictive performance while reducing model complexity. The network model also facilitates the exploration of feature interactions, providing a deeper understanding of the relationships driving birth defect outcomes.

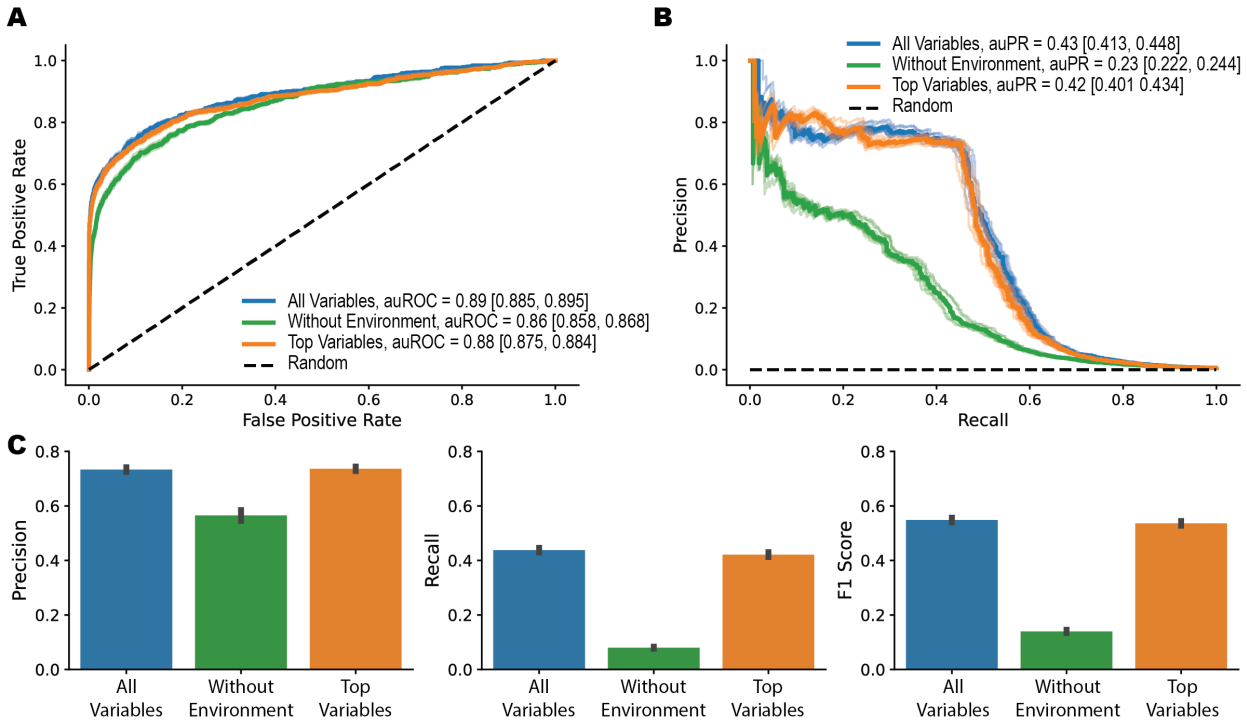

Figure 4.7: The area under the receiver operating curve (A), area under the precision-recall curve (B), and the precision, recall, and F1 score metrics (C) evaluated on the test set for XGBoost using three different feature sets: *All Variables*, *Without Environment*, and *Top Variables*. Lighter lines in (A) and (B) represent the 10-fold stratified test sets, while the thicker lines indicate the stratified mean, with confidence intervals for auROC and auPR provided in the legend. In (C), the bars represent the standard deviation across folds for precision, recall, and F1. The feature sets are defined as follows: *All Variables* includes all variables in the dataset, *Without Environment* excludes variables related to the physical and social environment, and *Top Variables* comprises the top 30 variables identified by the multilayer network model.

Feature Importance: Reduced Feature Space Improves Interpretability with Impaired Water Bodies as a Key Environmental Predictor

To identify the most predictive variables for birth defect outcomes, we computed SHAP (SHapley Additive exPlanations) values for both the *All Variables* and *Top Variables* feature sets. Global feature importance was determined by averaging SHAP values across all samples, while local feature importance was assessed using individual-specific SHAP values. This dual approach allowed us to capture both broad trends and individual-level insights.

The features contributing globally to model predictions for the *All Variables* feature set by variable subset is provided in Figure 4.8A with the top 20 individual features provided in Figure 4.8B. Utilizing the additive nature of SHAP values and normalizing by the number of features generate from each feature set, the physical and social environment overwhelmingly explains model predictions accounting for 60% of weighted SHAP explanations. The *Hearing Test Pass (Both Ears)* feature was the most important global feature contributing to birth defect outcomes followed by *Other Pregnancy Complications*, *Toxic Release from Facilities*, *APGAR Score (Min5)*, and *Ozone*. Interestingly, when looking at the true SHAP values for each individual for a local representation of feature importance (Figure 4.8C) *Impaired Water Bodies* has the maximum SHAP value with high values of *Impaired Water Bodies* having a positive impact on birth defect predictions. The second variable with the largest SHAP value is *Unemployment*, also coming from the physical and social environment subset.

To understand the nonlinear relationships between a feature and model output, we produced dependence plots that represent the SHAP values for a specific feature in terms of a Relative Risk (RR) ratio. Presented in Figure 4.8D-G are the relative risk for the top 4 continuous features with an exhaustive depiction of the top 20 in Figure C.1. The large SHAP values for *Impaired Water Bodies* is further demonstrated with large values of *Impaired Water Bodies* increasing the potential risk of birth defects ($RR > 1$). Interestingly, looking at the

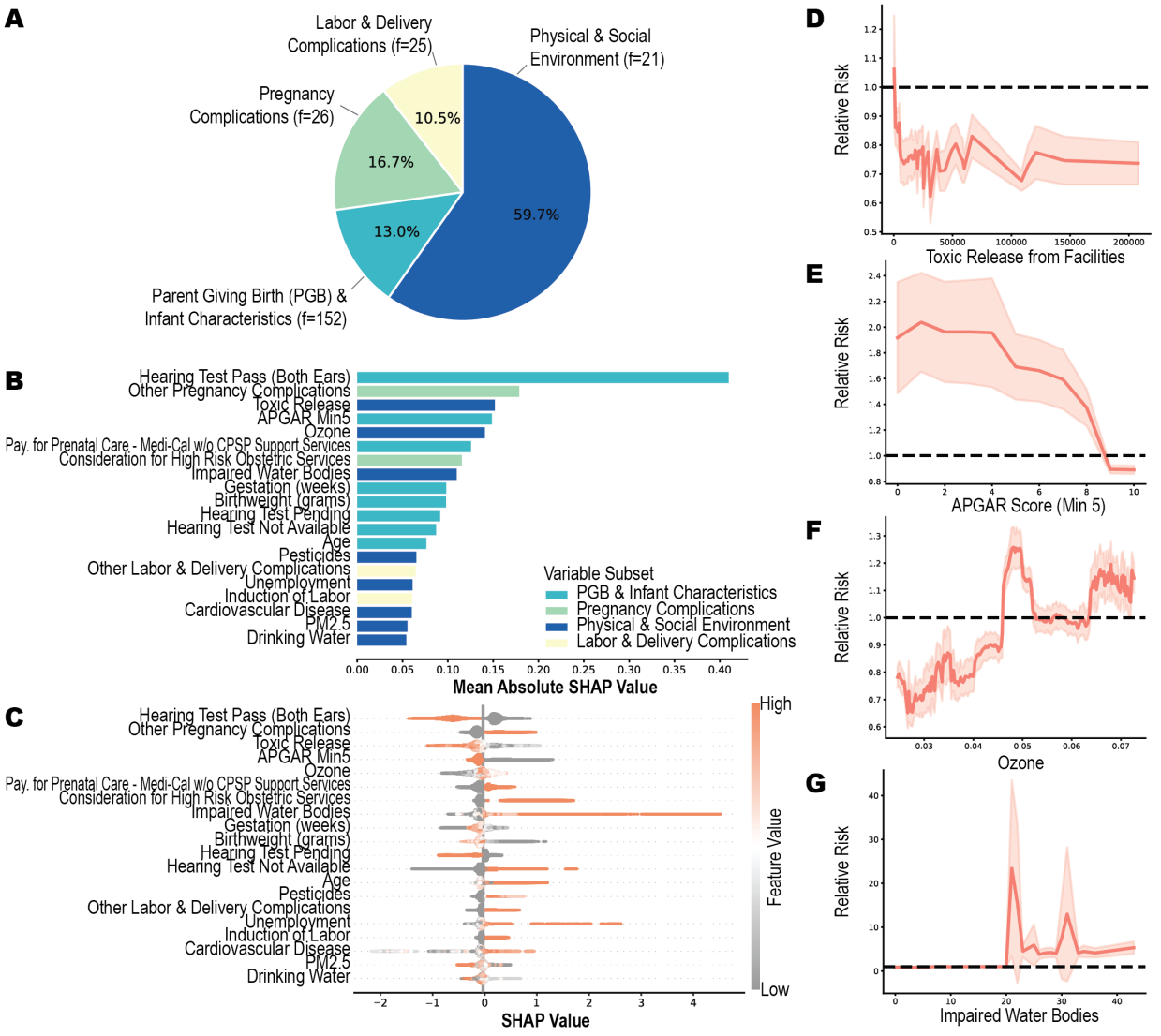

Figure 4.8: Feature importance and model interpretations using SHAP values generated with the *All Variables* feature set. (A) Global feature importance normalized to the number of features within each variable subset. SHAP values were summed for each variable subset and normalized to the number of features within the subset. The legend indicates the number of features (f) in each variable subset. Normalized contribution and non-transformed contribution is shown in Figure C.2 (B) The top 20 features ranked by mean absolute SHAP values. (C) Beeswarm plot of individual feature importances, where each dot represents the SHAP value of a feature for a specific observation within the model predictions. (D–G) Dependence plots for the top four continuous features contributing to model output, presented as risk ratios (RR). Shaded bands represent the standard deviation of the population per bin, and dashed lines indicate the baseline risk ratio ($RR = 1$). $RR > 1$ signifies a positive relationship with the model output, while $RR < 1$ signifies a negative relationship.

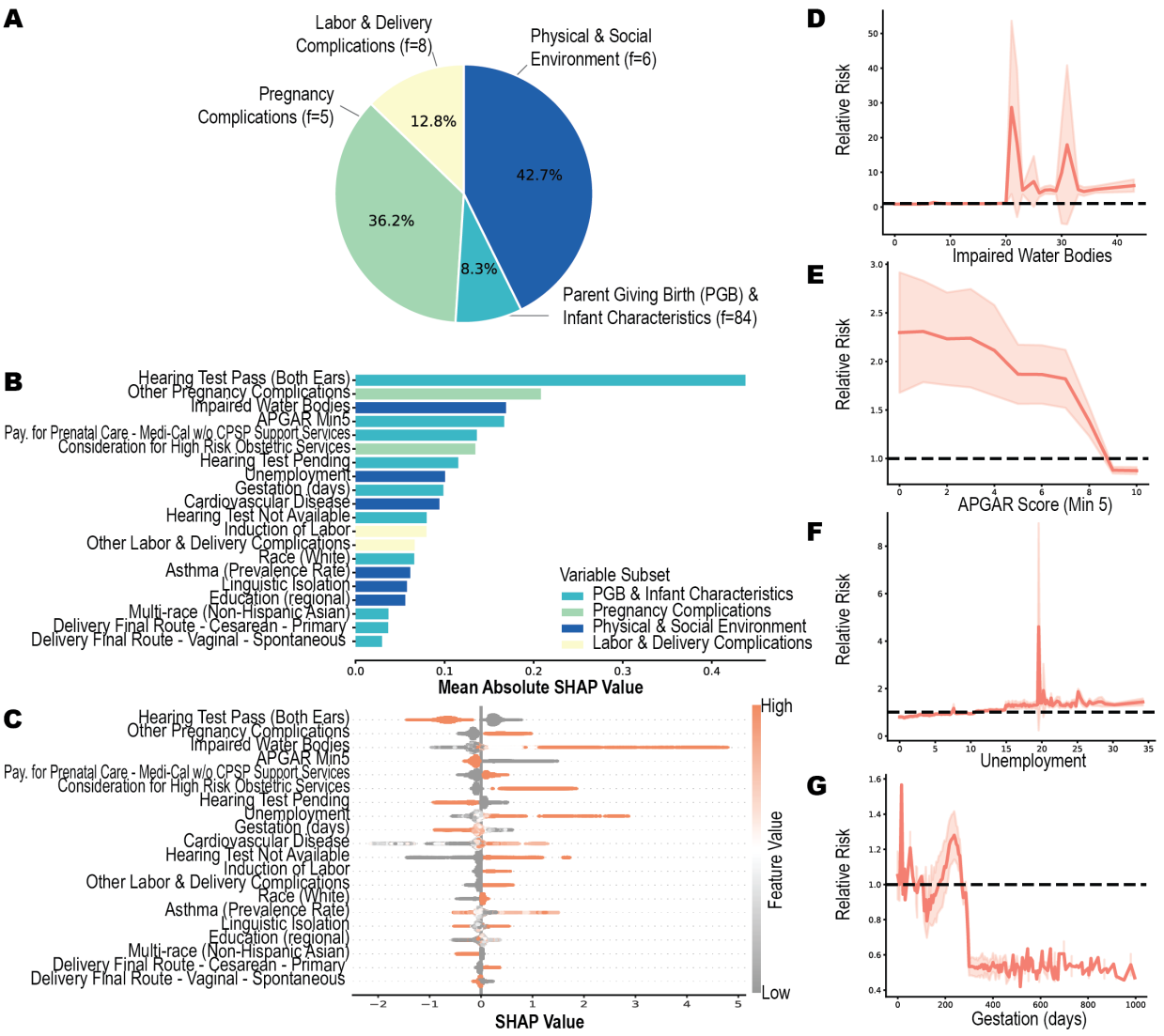

Figure 4.9: Feature importance and model interpretations using SHAP values generated with the *Top Variables* feature set. (A) Global feature importance normalized to the number of features within each variable subset. SHAP values were summed for each variable subset and normalized to the number of features within the subset. The legend indicates the number of features (f) in each variable subset. Normalized contribution and non-transformed contribution is shown in Figure C.4. (B) The top 20 features ranked by mean absolute SHAP values. (C) Beeswarm plot of individual feature importances, where each dot represents the SHAP value of a feature for a specific observation within the model predictions. (D–G) Dependence plots for the top four continuous features contributing to model output, presented as risk ratios (RR). Shaded bands represent the standard deviation of the population per bin, and dashed lines indicate the baseline risk ratio ($RR = 1$). $RR > 1$ signifies a positive relationship with the model output, while $RR < 1$ signifies a negative relationship.

Toxic Release from Facilities dependence plot, we see a near horizontal relationship between relative risk and *Toxic Release* with an $RR < 1$. This indicates that changes in *Toxic Release* values have little to no direct effect on the model's predicted birth defect risk. This suggests that the feature's observed importance, as shown by the SHAP values, is likely driven by its interactions or high correlations with other influential variables.

The reduced feature space informed by the multilayer network model provides a more streamlined approach to interpreting feature impact on model predictions. This reduction did not greatly alter prediction accuracy, suggesting that features excluded from the *Top Variables* feature set likely had minimal or no true impact on the model (AUROC reduction by 1%). Repeating the SHAP analysis for the *Top Variables* feature set yielded similar global insights, with the physical and social environment variable subset maintaining the highest contribution, accounting for 43% of the weighted SHAP values (Figure 4.9A). *Hearing Test Pass* and *Other Pregnancy Complications* remained the top two features by global average SHAP values (Figure 4.9B).

Notably, features like *Toxic Release* and *Ozone Levels*, which appeared in the top 20 in the *All Variables* feature set, were absent in the reduced feature set. This suggests that these variables may have served as proxies for other, more impactful variables, and were likely highly correlated with features identified as key by the multilayer network model. Interestingly, the two features with the highest positive impact on model predictions in the *All Variables* feature set continued to have the greatest contributions in the *Top Variables* feature set.

This result highlights the advantage of a reduced feature set in mitigating bias introduced by redundant features. For the top four continuous features, environmental factors such as *Impaired Water Bodies* and *Unemployment* were associated with an increased risk of birth defects ($RR > 1$). Conversely, features like *Gestation Time* showed a negative relationship, with shorter gestation times increasing the risk of birth defects. The relative risk for the top

20 features in the *Top Variables* feature set is included in Figure C.3.

4.4 Discussion

In this study we utilized an integrative computational modeling approach combining multilayer network modeling and machine learning to investigate the multifactorial relationship between birth defect outcomes and the parent giving birth comorbidities, socioeconomic factors, demographic information, and the physical and social environment the parent lives in. By incorporating data from two primary sources, live birth records form the CDPH and environmental data from the CalEPA, we were able to examine interactions across these domains on a population scale, enhancing interpretability and generalizability. Our findings reveal that the birth defects cohort exhibits increased variable co-occurrence across data domains, as demonstrated by higher inter-layer connectivity and clustering behavior in the multilayer network model. This underscores the multifactorial interplay of variables that may collectively contribute to birth defect incidence. The developed machine learning model exhibited strong predictive performance for birth defect outcomes (AUROC = 0.88) on the basis of a reduced feature space to influential variables identified through multilayer network analysis, improving both interpretability and efficiency.

Multilayer network density analysis revealed a significant increase in both inter- and intra-layer connections for the birth defects cohort, with the notable exception of intra-layer density in the Physical & Social Environment layer. This finding highlights a unique connectivity pattern in this domain, suggesting that variables from the physical and social environment may exert their influence through cross-domain interactions rather than internal cohesion. These results align with findings in environmental health research, which often highlight the compounded effects of factors such as proximity to pollutants and socioeconomic disparities on adverse outcomes [163, 164, 165]. Clustering analysis further supported this interpreta-

tion, identifying stronger co-occurrence patterns across data domains in the birth defects cohort, reflecting the potential compounding effects of correlated variables. Key variables driving connectivity in the birth defects cohort included hearing outcomes, other pregnancy complications, antibiotic use, multiracial identity, and labor and delivery complications. The prominence of these variables suggests they co-occur at higher rates with other network variables, highlights critical disparities in healthcare access within this cohort. For example, poor hearing outcomes, known to be associated with congenital abnormalities [166, 167], were prevalent, yet hearing tests were unavailable for over 60% of this cohort (Figure C.5). This discrepancy may reflect disparities in healthcare access, as government-supported health insurance covered prenatal and delivery care for more than 60% of the birth defects cohort (Figure C.5). These findings are further supported by studies demonstrating an association between hearing test failure and demographic factors such as public health insurance coverage, race, education, and geographic location [168, 169] while many high risk newborns have been shown to not receive screening at all [170]. Additionally, multiracial identity emerged as a top node, with approximately 70% of the birth defects cohort identified as Hispanic. These findings contrast with the without-birth-defects cohort, where variables such as payment methods for prenatal care and the number of children born alive dominated, reflecting broader population-level patterns rather than direct associations with adverse outcomes.

Our machine learning model achieved high predictive performance (AUROC = 0.88) for birth defect outcomes using a reduced feature space derived from the multilayer network analysis. Excluding the physical and social environment variables significantly reduced predictive accuracy, emphasizing their critical role in birth defect outcomes. To address challenges associated with class imbalance and overfitting in machine learning models, we implemented a stratified testing approach, class weighting, and hyperparameter optimization, ensuring robust performance. Additionally, we investigated tree-based machine learning models, such as the final XGBoost utilized here, have been shown to boost interpretability and be the most aligned with traditional clinical interpretation, emphasizing their applicability to the clinical

domain [171]. While our model has yet to be tested on an external test set, our testing leave-out method simulates model performance on an external dataset. Our best-performing model, XGBoost, achieved AUROCs of 0.89, 0.86, 0.88 and AUPRs of 0.43, 0.23, 0.42 for *All Variables*, *Without Environment*, and *Top Variables* feature spaces, respectively. Interestingly, the *Without Environment* feature space greatly lacked predictive power. The reduced feature space *Top Variables* demonstrated only a 1% decrease in performance metrics, highlighting the utility of our network-guided feature selection. An advantage of machine learning is its ability to capture nonlinear relationships between predictors and outcomes, which are often overlooked in traditional risk models [172]. In this approach, we were able to determine nonlinear risk ratios using SHAP values of feature importance. A key example of nonlinear behavior is the age of the mother giving birth for which low ages have increased risk, risk decreases as age increases, and risk increases again at later ages (Figure C.5). Notably, the hearing test category emerged as the top feature in the machine learning model (Figure 4.9) and the multilayer network (Table 4.3), with outcomes of "fail" or "pass" increasing and decreasing birth defect risk respectively (Figure 4.9C). Additionally, physical and social environment variables such as proximity to impaired water bodies and high unemployment rates had the largest local impacts on predictions, identifying vulnerable populations and geographies for targeted public health interventions.

Our approach of integrating multilayer network modeling and machine learning was able to uncover complex, multifactorial relationships while enhancing predictive accuracy and interpretability. Multilayer network modeling excels in identifying groups of highly correlated variables across diverse domains [173, 174], offering insights into their interrelationships and serving as a foundation for refining our machine learning model through feature reduction. This process mitigates the risk of skewed inference caused by multicollinearity, where models might disproportionately rely on one variable over others, leading to misleading conclusions [175]. For instance, in the *All Variable* feature space, *Gestation Time (weeks)* was associated with an unexpected increase in birth defect risk, contrary to established evidence that

shorter gestation times are more strongly linked to such outcomes [176]. This anomaly occurred because the model was primarily relying on *Gestation Time (days)*, where shorter durations indeed showed an increased risk in the *Top Variables* feature space. By utilizing the reduced feature space therefore removing correlated variables, the model becomes more interpretable and aligned with known risk factors, ensuring predictions are both robust and scientifically valid. In addition to enhancing interpretability, the multilayer network model provides a framework to infer the broader implications of variables identified as significant by the machine learning model. For example, *Impaired Water Bodies* emerged as one of the most influential variables, positively impacting birth defect risk predictions (Figure 4.9). Through the network model, we identified that this variable is highly correlated with other environmental and socioeconomic factors within the with-birth defects cohort, including *Pesticides* and *Unemployment*, which exhibited the largest effect sizes, as well as *Groundwater Threats*, *Toxic Release*, *Ozone*, *Linguistic Isolation*, *Cardiovascular Disease*, and *Solid Waste* (Figure 4.3). Although not all these variables were directly highlighted by the machine learning model, their strong correlation and co-occurrence with *Impaired Water Bodies* underscore the interconnected nature of the physical and social environment in shaping birth defect risks. This integration of network modeling and machine learning thus reveals potential confounding factors and complex interactions that may otherwise remain obscured. Moreover, the iterative interplay between these methodologies exemplifies their complementary strengths, which have been leveraged for other health related applications [177]. By narrowing the feature space to the most relevant variables identified through the network model, we maintained predictive performance while improving model efficiency and interpretability. This dual approach enables a more nuanced understanding of the multifactorial interactions underlying birth defect outcomes and offers a roadmap for digging deeper into machine learning interpretations, which is often overlooked [178]. Importantly, it highlights the compounded risks faced by vulnerable populations exposed to overlapping physical and social environmental hazards.

The multifactorial nature of birth defect outcomes, as revealed through our machine learning and multilayer network analyses, underscores that these outcomes are rarely attributable to a single cause. Instead, our approach enables the identification of key environmental, social, and health variables that are strongly correlated with or indicative of birth defects, offering actionable insights for policymakers. For example, *Impaired Water Bodies* emerged as a central variable influencing birth defect outcomes both in the multilayer network and machine learning model. This finding underscores the critical intersection of environmental and socioeconomic factors and aligns with evidence that environmental issues, such as water quality degradation, disproportionately impact lower-income communities and people of color. Addressing this variable through targeted policy initiatives, such as improving water quality in vulnerable areas, could mitigate risks while simultaneously addressing systemic factors like unemployment and healthcare access. This approach also aids in resource allocation by identifying high-risk demographic groups and geographic regions where interventions could have the most significant impact. For example, pairing water quality improvement efforts with initiatives addressing unemployment and social isolation could address the interconnected risks associated with birth defects more holistically. Such efforts are currently being implemented by the state of California, where the CalEnviroScreen program prioritizes funding for communities disproportionately impacted by environmental pollutants under the Senate Bill (SB) 535 [179]. CalEnviroScreen employs computational models to evaluate both environmental hazards and socioeconomic vulnerabilities, providing a data-driven assessment of pollution burden that informs regulatory and funding decisions [179, 180]. In contrast, the findings of this study extend this framework by specifically examining birth defect outcomes in relation to environment and socioeconomic factors. This targeted approach not only deepens our understanding of the multifactorial nature of birth defects but also offers an additional layer of insight for public health interventions. By pinpointing central variables and considering their broader correlations, this integrated strategy has the potential to inform preventive measures that are both effective and equitable, advancing public health

outcomes in a multifactorial context.

Numerous studies have documented the increased risk of birth defects associated with environmental factors [181, 182, 183, 4, 184]. While these studies often focus on univariate relationships, examining the effect of a single variable on specific birth defect outcomes, research by Peyvandi et al. underscores the multifactorial nature of birth defects, particularly congenital heart disease, by linking socioeconomic and environmental factors across California [19]. Their findings, which demonstrate an increased risk of birth defects associated with social deprivation, align with our results and incorporate a composite measure of social deprivation that aggregates multiple factors to account for confounders. Similarly, other studies have adopted nonlinear approaches using machine learning to uncover complex variable relationships and their impact on birth defect outcomes [185, 152]. Our study builds on this progress by integrating multilayer network modeling with machine learning, capturing relationships across multiple domains and leveraging network-derived insights to enhance predictive accuracy. This dual approach addresses a common limitation in traditional regression and machine learning frameworks: the tendency to overlook correlated or interdependent variables, which can obscure interpretations [186]. By incorporating multifactorial relationships, our methodology provides a more nuanced understanding of the drivers of birth defect outcomes, advancing the field beyond existing approaches.

Despite its strengths, our integrated modeling strategy has several limitations. First, the reliance on retrospective health records introduces potential biases, including the misclassification of outcomes and errors in the recording of birth monitoring codes [187, 188]. To mitigate these issues, we conducted rigorous data preprocessing, ensuring that observations labeled as *Normal* are excluded from any birth defect classifications. We also restricted our analysis to cases with clearly defined birth defect categories, excluding non-specific labels included within CCMBF records such as *Additional Abnormal Conditions* to enhance classification accuracy. Second, the use of the CalEnviroScreen tool, while comprehensive, is

sensitive to methodological changes and interpretative variability [189]. To address this, we supplemented our analyses with raw data from the CalEnviroScreen framework, enhancing transparency and reducing dependence on aggregated indices. Another challenge lies in the rarity of birth defect outcomes, which limits the availability of robust validation datasets. This is a common issue in machine learning studies focused on rare events [190, 191]. We addressed it by employing techniques to manage class imbalance, including thorough validation with leave-out test sets. Nonetheless, further validation with external datasets is necessary to confirm the generalizability of our findings. Finally, our study focuses on data from California, a state with diverse socioeconomic and demographic profiles. While this diversity enhances the relevance of our findings, the applicability of our predictive model to other populations remains untested. Future studies should explore the generalizability of our model in different geographic and demographic contexts.

In conclusion, our findings demonstrate that birth defect outcomes are inherently multifactorial, with multilayer network modeling enhancing understanding and machine learning achieving high predictive accuracy. This integrated approach not only identifies key variables driving adverse outcomes but also provides a foundation for prioritizing interventions that address the multifaceted risks faced by vulnerable populations. These results have significant implications for public health and environmental mediation strategies, offering a pathway to improve health outcomes through targeted, evidence-based policies.

Chapter 5

Conclusion

In this dissertation, I have developed and presented a suite of integrative computational tools to assess the impact of environmental exposures during development. By combining advanced methodologies, ranging from enhanced transcriptomic analysis to multilayer network modeling and machine learning, this work addresses key challenges in developmental toxicology at multiple biological scales. The tools and approaches presented here not only bridge gaps between molecular and population-level analyses but also offer scalable frameworks for future research.

Throughout the dissertation, three major contributions have been highlighted:

- **Chapter 2:** I introduced danRerLib, a Python package tailored for zebrafish transcriptomics and functional enrichment analysis. By leveraging orthology mapping between zebrafish and human gene annotations, danRerLib overcomes the limitations imposed by incomplete zebrafish pathway annotations. This tool enhances our ability to identify biologically relevant pathways affected by environmental exposures, serving as a foundational resource for subsequent analyses.

- **Chapter 3:** Building on the capabilities of danRerLib, I developed a computational framework that integrates network modeling and machine learning to predict chemical toxicity at the organ level, focusing on the pancreas. This work demonstrates that whole-embryo zebrafish transcriptomic data can be repurposed to extract organ-specific insights. The combined analysis not only identifies key driver genes and potential biomarkers of pancreatic toxicity but also establishes a proof-of-concept model for organ-targeted toxicity prediction.

- **Chapter 4:** The methodology was further extended to the population level by applying multilayer network modeling and machine learning to a comprehensive dataset from California. This analysis elucidated the complex interactions between environmental exposures, demographic factors, and birth defect outcomes. The integrative approach achieved strong predictive performance and highlighted critical predictors, such as impaired water bodies, thereby reinforcing the multifactorial nature of birth defect etiology.

5.1 Future Work

The projects presented in this dissertation have been designed with scalability in mind, and several avenues for future work are evident.

While Chapter 2 focuses on functional enrichment analysis using danRerLib, future enhancements could include integrating upstream bioinformatics processes such as differential expression analysis, read alignment, and gene counting specifically optimized for zebrafish datasets. Additionally, a server-based graphical user interface (GUI) would increase access to danRerLib by eliminating the need for programming expertise, thus broadening its adoption among researchers with limited computational backgrounds.

Chapter 3 established a foundational toxicity prediction model for the pancreas. An important next step is extending this framework to other organ systems. One promising direction involves developing a multilayer network model where each layer represents a different organ. Challenges such as establishing robust intra-layer links between organ pathways, potentially via gene overlap, and reducing the network's dimensionality for improved clustering and prediction should be addressed. Moreover, as new chemicals are characterized, they can be integrated into the chemical space, thereby refining clustering outcomes and enhancing the predictive power of the model. Developing a user-friendly interface to input chemical properties and receive toxicity predictions for multiple organs would further increase the model's utility.

In Chapter 4, the application of multilayer network modeling and machine learning to California's population data yielded promising results. Future work should focus on validating these models on external datasets and exploring additional machine learning approaches that can scale with increasing data volume. As the dataset expands, novel models may further improve the predictive performance and interpretability of key risk factors associated with birth defects.

A final important direction is the construction of a comprehensive Adverse Outcome Pathway (AOP) that links molecular events to population-level outcomes. This work is currently limited by the disparity in chemical types between the zebrafish transcriptomics dataset and those monitored through programs like CalEnviroScreen. Future research should strive to map identifiable chemicals across datasets and generate robust zebrafish transcriptomic profiles for these chemicals. Such efforts would facilitate direct linkage of molecular toxicity predictions (Chapter 3) to observed birth defect outcomes (Chapter 4), while accounting for the multifactorial influence of demographic and socioeconomic variables.

5.2 Concluding Remarks

The long-term goal of this research is to advance our understanding of how environmental contaminants affect embryonic development, a critical step toward mitigating adverse health outcomes. In summary, this dissertation has provided novel computational frameworks that integrate transcriptomic analysis, network modeling, and machine learning to address the challenges of developmental toxicity. The tools developed here, including danRerLib and the organ-specific as well as population-level models, represent significant progress toward predictive toxicology. They not only fill existing gaps in our current methodologies but also offer scalable, adaptable approaches that can guide future research, inform public health policies, and ultimately contribute to the prevention of environmentally induced developmental disorders.

Bibliography

[1] Ashley V. Schwartz, Karilyn E. Sant, and Uduak Z. George. danrerlib: a python package for zebrafish transcriptomics. *Bioinformatics Advances*, 4(1), 2024.

[2] Ashley V. Schwartz, Karilyn E. Sant, and Uduak Z. George. Integrating network analysis and machine learning to elucidate chemical-induced pancreatic toxicity in zebrafish embryos. *Toxicological Sciences*, 2025.

[3] Ashley E. Larsen, Steven D. Gaines, and Olivier Deschênes. Agricultural pesticide use and adverse birth outcomes in the san joaquin valley of california. *Nature Communications*, 8(1):302, 2017.

[4] Beate Ritz, Fei Yu, Scott Fruin, Guadalupe Chapa, Gary M. Shaw, and John A. Harris. Ambient air pollution and risk of birth defects in southern california. *American Journal of Epidemiology*, 155(1):17–25, 2002.

[5] Suzan L. Carmichael, Wei Yang, Eric Roberts, Susan E. Kegley, Timothy J. Brown, Paul B. English, Edward J. Lammer, and Gary M. Shaw. Residential agricultural pesticide exposures and risks of selected birth defects among offspring in the san joaquin valley of california. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 106(1):27–35, 2016.

[6] Linda M. Frazier. Reproductive disorders associated with pesticide exposure. *Journal of Agromedicine*, 12(1):27–37, 2007.

[7] Akash Sabarwal, Kunal Kumar, and Rana P. Singh. Hazardous effects of chemical pesticides on human health–cancer and other associated disorders. *Environmental Toxicology and Pharmacology*, 63:103–114, 2018.

[8] D Rice and S Barone. Critical periods of vulnerability for the developing nervous system: evidence from humans and animal models. *Environmental Health Perspectives*, 108(suppl 3):511–533, 2000.

[9] H. T. Mocelin, G. B. Fischer, and A. Bush. Adverse early-life environmental exposures and their repercussions on adult respiratory health. *J Pediatr (Rio J)*, 98 Suppl 1(Suppl 1):S86–s95, 2022.

[10] Ines Amine, Alicia Guillien, Claire Philippat, Augusto Anguita-Ruiz, Maribel Casas, Montserrat de Castro, Audrius Dedele, Judith Garcia-Aymerich, Berit Granum, Regina

Grazuleviciene, Barbara Heude, Line Småstuen Haug, Jordi Julvez, Mónica López-Vicente, Léa Maitre, Rosemary McEachan, Mark Nieuwenhuijsen, Nikos Stratakis, Marina Vafeiadi, John Wright, Tiffany Yang, Wen Lun Yuan, Xavier Basagaña, Rémy Slama, Martine Vrijheid, and Valérie Siroux. Environmental exposures in early-life and general health in childhood. *Environmental Health*, 22(1):53, 2023.

[11] Tim D. Williams, Leda Mirbahai, and J. Kevin Chipman. The toxicological application of transcriptomics and epigenomics in zebrafish and other teleosts. *Briefings in Functional Genomics*, 13(2):157–171, 2014.

[12] Carlos Pardo-Martin, Tsung-Yao Chang, Bryan Kyo Koo, Cody L. Gilleland, Steven C. Wasserman, and Mehmet Fatih Yanik. High-throughput in vivo vertebrate screening. *Nature Methods*, 7(8):634–636, 2010.

[13] David J. Dix, Keith A. Houck, Matthew T. Martin, Ann M. Richard, R. Woodrow Setzer, and Robert J. Kavlock. The toxcast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences*, 95(1):5–12, 2007.

[14] Brian A. Link and Sean G. Megason. *Zebrafish as a Model for Development*, pages 103–112. Humana Press, Totowa, NJ, 2008.

[15] Julia F. Gilmartin-Thomas, Danny Liew, and Ingrid Hopper. Observational studies and their utility for practice. *Aust Prescr*, 41(3):82–85, 2018.

[16] Lixin Yang, Nga Yu Ho, Rüdiger Alshut, Jessica Legradi, Carsten Weiss, Markus Reischl, Ralf Mikut, Urban Liebel, Ferenc Müller, and Uwe Strähle. Zebrafish embryos as models for embryotoxic and teratological effects of chemicals. *Reproductive Toxicology*, 28(2):245–253, 2009.

[17] Charles B. Kimmel, William W. Ballard, Seth R. Kimmel, Bonnie Ullmann, and Thomas F. Schilling. Stages of embryonic development of the zebrafish. *Developmental Dynamics*, 203(3):253–310, 1995.

[18] Prarthana Shankar, Ryan S. McClure, Katrina M. Waters, and Robyn L. Tanguay. Gene co-expression network analysis in zebrafish reveals chemical class specific modules. *BMC Genomics*, 22(1):658, 2021.

[19] Shabnam Peyvandi, Rebecca J. Baer, Christina D. Chambers, Mary E. Norton, Satish Rajagopal, Kelli K. Ryckman, Anita Moon-Grady, Laura L. Jelliffe-Pawlowski, and Martina A. Steurer. Environmental and socioeconomic factors influence the live-born incidence of congenital heart disease: A population-based study in california. *J. Am. Heart Assoc.*, 9(8):e015255, 2020.

[20] Emanuel Alcala, Paul Brown, John A. Capitman, Mariaelena Gonzalez, and Ricardo Cisneros. Cumulative impact of environmental pollution and population vulnerability on pediatric asthma hospitalizations: A multilevel analysis of calenviroscreen. *Int. J. Environ. Res. Public Health*, 16(15):2683, 2019.

[21] Gretchen Bandoli, Rebecca J. Baer, Mallory Owen, Elizabeth Kiernan, Laura Jelliffe-Pawlowski, Stephen Kingsmore, and Christina D. Chambers. Maternal, infant, and environmental risk factors for sudden unexpected infant deaths: results from a large, administrative cohort. *The Journal of Maternal-Fetal & Neonatal Medicine*, 35(25):8998–9005, 2022.

[22] Lara Cushing, John Faust, Laura Meehan August, Rose Cendak, Walker Wieland, and George Alexeeff. Racial/ethnic disparities in cumulative environmental health impacts in california: Evidence from a statewide environmental justice screening tool (calenviroscreen 1.1). *American Journal of Public Health*, 105(11):2341–2348, 2015.

[23] Saleha Khanum, Zohir Chowdhury, and Karilyn E. Sant. Association between particulate matter air pollution and heart attacks in san diego county. *Journal of the Air & Waste Management Association*, 71(12):1585–1594, 2021.

[24] Jing Chen, Yuan Yan Tang, Bin Fang, and Chang Guo. In silico prediction of toxic action mechanisms of phenols for imbalanced data with random forest learner. *Journal of Molecular Graphics and Modelling*, 35:21–27, 2012.

[25] Y. G. Chushak, H. W. Shows, J. M. Gearhart, and H. A. Pangburn. In silico identification of protein targets for chemical neurotoxins using toxcast in vitro data and read-across within the qsar toolbox. *Toxicol Res (Camb)*, 7(3):423–431, 2018.

[26] Gongqing Dong, Nan Wang, Ting Xu, Jingyu Liang, Ruxia Qiao, Daqiang Yin, and Sijie Lin. Deep learning-enabled morphometric analysis for toxicity screening using zebrafish larvae. *Environmental Science & Technology*, 2023.

[27] Xiao Gou, Cong Ma, Huimin Ji, Lu Yan, Pingping Wang, Zhihao Wang, Yishan Lin, Nivedita Chatterjee, Hongxia Yu, and Xiaowei Zhang. Prediction of zebrafish embryonic developmental toxicity by integrating omics with adverse outcome pathway. *Journal of Hazardous Materials*, 448:130958, 2023.

[28] Adrian J. Green, Martin J. Mohlenkamp, Jhuma Das, Meenal Chaudhari, Lisa Truong, Robyn L. Tanguay, and David M. Reif. Leveraging high-throughput screening data, deep neural networks, and conditional generative adversarial networks to advance predictive toxicology. *PLoS Comput Biol*, 17(7):e1009135, 2021.

[29] Jennifer Hemmerich and Gerhard F. Ecker. In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. *WIREs Computational Molecular Science*, 10(4):e1475, 2020.

[30] K. T. Rim. In silico prediction of toxicity and its applications for chemicals at work. *Toxicol Environ Health Sci*, 12(3):191–202, 2020.

[31] Leihong Wu, Ruili Huang, Igor V. Tetko, Zhonghua Xia, Joshua Xu, and Weida Tong. Trade-off predictivity and explainability for machine-learning powered predictive toxicology: An in-depth investigation with tox21 data sets. *Chemical Research in Toxicology*, 34(2):541–549, 2021.

[32] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[33] Courtney Roper and Robert L. Tanguay. *Chapter 12 - Zebrafish as a Model for Developmental Biology and Toxicology*, pages 143–151. Academic Press, 2018.

[34] Kerstin Howe, Matthew D. Clark, Carlos F. Torroja, James Torrance, Camille Berthelot, Matthieu Muffato, John E. Collins, Sean Humphray, Karen McLaren, Lucy Matthews, Stuart McLaren, Ian Sealy, Mario Caccamo, Carol Churcher, Carol Scott, Jeffrey C. Barrett, Romke Koch, Gerd-Jörg Rauch, Simon White, William Chow, Britt Kilian, Leonor T. Quintais, José A. Guerra-Assunção, Yi Zhou, Yong Gu, Jennifer Yen, Jan-Hinnerk Vogel, Tina Eyre, Seth Redmond, Ruby Banerjee, Jianxiang Chi, Beiyuan Fu, Elizabeth Langley, Sean F. Maguire, Gavin K. Laird, David Lloyd, Emma Kenyon, Sarah Donaldson, Harminder Sehra, Jeff Almeida-King, Jane Loveland, Stephen Trevanion, Matt Jones, Mike Quail, Dave Willey, Adrienne Hunt, John Burton, Sarah Sims, Kirsten McLay, Bob Plumb, Joy Davis, Chris Clee, Karen Oliver, Richard Clark, Clare Riddle, David Elliott, Glen Threadgold, Glenn Harden, Darren Ware, Sharmin Begum, Beverley Mortimore, Giselle Kerry, Paul Heath, Benjamin Phillimore, Alan Tracey, Nicole Corby, Matthew Dunn, Christopher Johnson, Jonathan Wood, Susan Clark, Sarah Pelan, Guy Griffiths, Michelle Smith, Rebecca Glithero, Philip Howden, Nicholas Barker, Christine Lloyd, Christopher Stevens, Joanna Harley, Karen Holt, Georgios Panagiotidis, Jamieson Lovell, Helen Beasley, Carl Henderson, Daria Gordon, Katherine Auger, Deborah Wright, Joanna Collins, Claire Raisen, Lauren Dyer, Kenric Leung, Lauren Robertson, Kirsty Ambridge, Daniel Leongamornlert, Sarah McGuire, Ruth Gilderthorp, Coline Griffiths, Deepa Manthravadi, Sarah Nichol, Gary Barker, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446):498–503, 2013.

[35] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[36] Alex Haimbaugh, Chia-Chen Wu, Camille Akemann, Danielle N. Meyer, Mackenzie Connell, Mohammad Abdi, Aicha Khalaf, Destiny Johnson, and Tracie R. Baker. Multi- and transgenerational effects of developmental exposure to environmental levels of pfas and pfas mixture in zebrafish (danio rerio). *Toxics*, 10(6):334, 2022.

[37] Julian Navarrete, Peyton Wilson, Nicholas Allsing, Chandi Gordon, Rachel Margolis, Ashley V. Schwartz, Christine Cho, Brynn Rogowski, Jennifer Topps, Uduak Z. George, and Karilyn E. Sant. The ecotoxicological contaminant tris(4-chlorophenyl)methanol (tcpmoh) impacts embryonic development in zebrafish (danio rerio). *Aquatic Toxicology*, 235:105815, 2021.

[38] Edward J. Perkins, Kimberly T. To, Lindsey St Mary, Charles H. Laber, Anthony J. Bednar, Lisa Truong, Robyn L. Tanguay, and Natalia Garcia-Reyero. Developmental, behavioral and transcriptomic changes in zebrafish embryos after smoke dye exposure. *Toxics*, 10(5), 2022.

[39] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*, 51(D1):D587–d592, 2023.

[40] The Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Res*, 49(D1):D325–d334, 2021.

[41] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, 39(suppl_1):D52–D57, 2011.

[42] Fergal J Martin, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, Simarpreet Kaur Bhurji, Alexandra Bignell, Sanjay Boddu, Paulo R Branco Lins, Lucy Brooks, Shashank Budhanuru Ramaraju, Mehrnaz Charkhchi, Alexander Cockburn, Luca Da Rin Fiorretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Gurpreet S Ghattaoraya, Jose Gonzalez Martinez, Cristi Guijarro, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Mike Kay, Vinay Kaykala, Tuan Le, Diana Lemos, Diego Marques-Coelho, José Carlos Marugán, Gabriela Alejandra Merino, Louisse Paola Mirabueno, Aleena Mushtaq, Syed Nakib Hossain, Denye N Ogeh, Manoj Pandian Sakthivel, Anne Parker, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, José G Pérez-Silva, Ahamed Imran Abdul Salam, Nuno Saraiva-Agostinho, Helen Schuilenburg, Dan Sheppard, Swati Sinha, Botond Sipos, William Stark, Emily Steed, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Likhitha Surapaneni, Kyösti Sutinen, Michal Szpak, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Elizabeth Wass, Natalie Willhoft, Jamie Allen, Jorge Alvarez-Jarreta, Marc Chakiachvili, Bethany Flint, Stefano Giorgetti, Leanne Haggerty, Garth R Ilsley, Jane E Loveland, Benjamin Moore, Jonathan M Mudge, John Tate, David Thybert, Stephen J Trevanion, Andrea Winterbottom, Adam Frankish, Sarah E Hunt, Magali Ruffier, Fiona Cunningham, Sarah Dyer, Robert D Finn, Kevin L Howe, Peter W Harrison, Andrew D Yates, and Paul Flicek. Ensembl 2023. *Nucleic Acids Research*, 51(D1):D933–D941, 2022.

[43] Yvonne M Bradford, Ceri E Van Slyke, Leyla Ruzicka, Amy Singer, Anne Eagle, David Fashena, Douglas G Howe, Ken Frazer, Ryan Martin, Holly Paddock, Christian Pich, Sridhar Ramachandran, and Monte Westerfield. Zebrafish information network, the knowledgebase for danio rerio research. *Genetics*, 220(4), 2022.

[44] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[45] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[46] Chee Lee, Snehal Patil, and Maureen A. Sartor. Rna-enrich: a cut-off free functional enrichment testing method for rna-seq with improved detection power. *Bioinformatics*, 32(7):1100–1102, 2015.

[47] Maureen A. Sartor, George D. Leikauf, and Mario Medvedovic. Lrpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–7, 2009.

[48] Brad T. Sherman, Ming Hao, Ju Qiu, Xialoi Jiao, Michael W. Baseler, H. C. Lane, Tomozumi Imamichi, and Weizhong Chang. David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*, 50(W1):W216–w221, 2022.

[49] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[50] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, 2023.

[51] Edward Y. Chen, Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, 2013.

[52] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicholas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*, 44(W1):W90–7, 2016.

[53] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1), 2022.

[54] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021.

[55] Guangchuag Yu, Li-Gen G. Wang, Yanyan Han, and Qing Y. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics*, 16(5):284–7, 2012.

[56] D. V. Klopfenstein, Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, Jeffrey M. Yunes, Olga Botvinnik, Mark Weigel, Will Dampier, Christophe Dessimoz, Patrick Flick, and Haibao Tang. Goatools: A python library for gene ontology analyses. *Scientific Reports*, 8(1):10872, 2018.

[57] Kuang-Hung Pan, Chih-Jian Lih, and Stanley N. Cohen. Effects of threshold choice on biological conclusions reached during analysis of gene expression by dna microarrays. *Proceedings of the National Academy of Sciences*, 102(25):8961–8965, 2005.

[58] Richard S. Judson, Keith A. Houck, Robert J. Kavlock, Thomas B. Knudsen, Matthew T. Martin, Holly M. Mortensen, David M. Reif, Daniel M. Rotroff, Imran Shah, Ann M. Richard, and David J. Dix. In vitro screening of environmental chemicals for targeted testing prioritization: The toxcast project. *Environmental Health Perspectives*, 118(4):485–492, 2010.

[59] Robert Kavlock, Kelly Chandler, Keith Houck, Sid Hunter, Richard Judson, Nicole Kleinstreuer, Thomas Knudsen, Matt Martin, Stephanie Padilla, David Reif, Ann Richard, Daniel Rotroff, Nisha Sipes, and David Dix. Update on epa's toxcast program: Providing high throughput decision support tools for chemical risk management. *Chemical Research in Toxicology*, 25(7):1287–1302, 2012.

[60] Raymond R. Tice, Christopher P. Austin, Robert J. Kavlock, and John R. Bucher. Improving the human hazard characterization of chemicals: A tox21 update. *Environmental Health Perspectives*, 121(7):756–765, 2013.

[61] Katherine K Coady, Ronald C Biever, Nancy D Denslow, Melanie Gross, Patrick D Guiney, Henrik Holbech, Natalie K Karouna-Renier, Ioanna Katsiadaki, Hank Krueger, Steven L Levine, Gerd Maack, Mike Williams, Jeffrey C Wolf, and Gerald T Ankley. Current limitations and recommendations to improve testing for the environmental assessment of endocrine active substances. *Integrated Environmental Assessment and Management*, 13(2):302–316, 2017.

[62] Lusa Truong, Sean M. Bugel, Anna Chlebowski, Crystal Y. Usenko, Michael T. Simonich, Staci L. Simonich, and Robyn L. Tanguay. Optimizing multi-dimensional high throughput screening using zebrafish. *Reprod Toxicol*, 65:139–147, 2016.

[63] Natascia Tiso, Enrico Moro, and Francesco Argenton. Zebrafish pancreas development. *Mol Cell Endocrinol*, 312(1-2):24–30, 2009.

[64] Haydee M. Jacobs, Karilyn E. Sant, Aviraj Basnet, Larissa M. Williams, Jennifer B. Moss, and Alicia R. Timme-Laragy. Embryonic exposure to mono(2-ethylhexyl) phthalate (mehp) disrupts pancreatic organogenesis in zebrafish (danio rerio). *Chemosphere*, 195:498–507, 2018.

[65] Karilyn E. Sant, Kate Annunziato, Sarah Conlin, Gregory Teicher, Phoebe Chen, Olivia Venezia, Gerald B. Downes, Yeonhwa Park, and Alicia R. Timme-Laragy. Developmental exposures to perfluorooctanesulfonic acid (pfos) impact embryonic nutrition, pancreatic morphology, and adiposity in the zebrafish, danio rerio. *Environmental Pollution*, 275:116644, 2021.

[66] Karilyn E. Sant, Haydee M. Jacobs, Katrina A. Borofski, Jennifer B. Moss, and Alicia R. Timme-Laragy. Embryonic exposures to perfluorooctanesulfonic acid (pfos) disrupt pancreatic organogenesis in the zebrafish, danio rerio. *Environmental Pollution*, 220:807–817, 2017.

[67] Karilyn E. Sant, Haydee M. Jacobs, Jiali Xu, Katrina A. Borofski, Larry G. Moss, Jennifer B. Moss, and Alicia R. Timme-Laragy. Assessment of toxicological perturbations and variants of pancreatic islet development in the zebrafish model. *Toxics*, 4(3):20, 2016.

[68] Karilyn E. Sant, Olivia L. Venezia, Paul P. Sinno, and Alicia R. Timme-Laragy. Perfluorobutanesulfonic acid disrupts pancreatic organogenesis and regulation of lipid metabolism in the zebrafish, danio rerio. *Toxicol Sci*, 167(1):258–268, 2019.

[69] Alicia R. Timme-Laragy, Karilyn E. Sant, Michelle E. Rousseau, and Philip J. diIorio. Deviant development of pancreatic beta cells from embryonic exposure to pcb-126 in zebrafish. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 178:25–32, 2015.

[70] Peyton W. Wilson, Christine Cho, Nicholas Allsing, Saleha Khanum, Pria Bose, Ava Grubschmidt, and Karilyn E. Sant. Tris(4-chlorophenyl)methane and tris(4-chlorophenyl)methanol disrupt pancreatic organogenesis and gene expression in zebrafish embryos. *Birth Defects Research*, 115(4):458–473, 2023.

[71] Mary D. Kinkel and Victoria E. Prince. On the diabetic menu: zebrafish as a model for pancreas development and function. *Bioessays*, 31(2):139–52, 2009.

[72] Hiroki Matsuda. Zebrafish as a model for studying functional pancreatic β cells development and regeneration. *Development, Growth & Differentiation*, 60(6):393–399, 2018.

[73] Zacary Zamora, Susanna Wang, Yen-Wei W. Chen, Graciel Diamante, and Xia Yang. Systematic transcriptome-wide meta-analysis across endocrine disrupting chemicals reveals shared and unique liver pathways, gene networks, and disease associations. *Environ Int*, 183:108339, 2024.

[74] Van A. Huynh-Thu, Alexandre Irrthum, Lous Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5(9), 2010.

[75] Ashley V. Schwartz, Karilyn E. Sant, and Uduak Z. George. Development of a dynamic network model to identify temporal patterns of structural malformations in zebrafish embryos exposed to a model toxicant, tris(4-chlorophenyl)methanol. *Journal of Xenobiotics*, 13(2):284–297, 2023.

[76] Heather L. Ciallella, Daniel P. Russo, Lauren M. Aleksunes, Fabian A. Grimm, and Hao Zhu. Revealing adverse outcome pathways from public high-throughput screening data to evaluate new toxicants by a knowledge-based deep neural network approach. *Environmental Science & Technology*, 55(15):10875–10887, 2021.

[77] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L.

Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Res*, 41(Database issue):D991–5, 2013.

[78] Andreas Schüttler, Kristin Reiche, Rolf Altenburger, and Wibke Busch. The transcriptome of the zebrafish embryo after chemical exposure: A meta-analysis. *Toxicological Sciences*, 157(2):291–304, 2017.

[79] U.S. Environmental Protection Agency. Comptox chemicals dashboard, 2024.

[80] Antony J. Williams, Christopher M. Grulke, Jeff Edwards, Andrew D. McEachran, Kamel Mansouri, Nancy C. Baker, Grace Patlewicz, Imran Shah, John F. Wambaugh, Richard S. Judson, and Ann M. Richard. The comptox chemistry dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics*, 9(1):61, 2017.

[81] Monika A. Roy, Karilyn E. Sant, Olivia L. Venezia, Alix B. Shipman, Stephen D. McCormick, Panithi Saktrakulkla, Keri C. Hornbuckle, and Alicia R. Timme-Laragy. The emerging contaminant 3,3'-dichlorobiphenyl (pcb-11) impedes ahr activation and cyp1a activity to modify embryotoxicity of ahr ligands in the zebrafish embryo model (danio rerio). *Environ Pollut*, 254(Pt A):113027, 2019.

[82] Maria Christou, Thomas W. K. Fraser, Vidar Berg, Erik Ropstad, and Jorke H. Kamstra. Calcium signaling as a possible mechanism behind increased locomotor response in zebrafish larvae exposed to a human relevant persistent organic pollutant mixture or pfos. *Environ Res*, 187:109702, 2020.

[83] Jing Huang, Qiyu Wang, Shuai Liu, Miao Zhang, Yu Liu, Liwei Sun, Yongming Wu, and Wenqing Tu. Crosstalk between histological alterations, oxidative stress and immune aberrations of the emerging pfos alternative obs in developing zebrafish. *Science of The Total Environment*, 774:145443, 2021.

[84] Zacharuas Pandelides, Neelakanteswar Aluru, Cammi Thornton, Haley E. Watts, and Kristine L. Willett. Transcriptomic changes and the roles of cannabinoid receptors and pparγ in developmental toxicities following exposure to δ9-tetrahydrocannabinol and cannabidiol. *Toxicol Sci*, 182(1):44–59, 2021.

[85] Rubén Martínez, Anna Esteve-Codina, Laia Herrero-Nogareda, Elena Ortiz-Villanueva, Carlos Barata, Romà Tauler, Demetrio Raldúa, Benjamin Piña, and Laia Navarro-Martín. Dose-dependent transcriptomic responses of zebrafish eleutheroembryos to bisphenol a. *Environmental Pollution*, 243:988–997, 2018.

[86] Rubén Martínez, Laia Navarro-Martín, Chiara Luccarelli, Anna E. Codina, Demetrio Raldúa, Carlos Barata, Romà Tauler, and Benjamin Piña. Unravelling the mechanisms of pfos toxicity by combining morphological and transcriptomic analyses in zebrafish embryos. *Science of The Total Environment*, 674:462–471, 2019.

[87] Rubén Martínez, Anna E. Codina, Carlos Barata, Romà Tauler, Benjamin Piña, and Laia Navarro-Martín. Transcriptomic effects of tributyltin (tbt) in zebrafish eleutheroembryos. a functional benchmark dose analysis. *Journal of Hazardous Materials*, 398:122881, 2020.

[88] Adriaan D. Bastiaan Vliegenthart, Chunmin Wei, Charlotte Buckley, Cécile Berends, Carmelita M. J. de Potter, Sarah Schneemann, Jorge Del Pozo, Carl Tucker, John J. Mullins, David J. Webb, and James W. Dear. Characterization of triptolide-induced hepatotoxicity by imaging and transcriptomics in a novel zebrafish model. *Toxicological Sciences*, 159(2):380–391, 2017.

[89] Cong Minh Tran, Jin-Sung Ra, Dong Young Rhyu, and Ki-Tae Kim. Transcriptome analysis reveals differences in developmental neurotoxicity mechanism of methyl-, ethyl-, and propyl- parabens in zebrafish embryos. *Ecotoxicology and Environmental Safety*, 268:115704, 2023.

[90] Neelakanteswar Aluru, Keegan S Krick, Adriane M McDonald, and Sibel I Karchner. Developmental exposure to pcb153 (2,2',4,4',5,5'-hexachlorobiphenyl) alters circadian rhythms and the expression of clock and metabolic genes. *Toxicological Sciences*, 173(1):41–52, 2019.

[91] Neelakanteswar Aluru, Sibel I Karchner, and Lilah Glazer. Early life exposure to low levels of ahr agonist pcb126 (3,3',4,4',5-pentachlorobiphenyl) reprograms gene expression in adult brain. *Toxicological Sciences*, 160(2):386–397, 2017.

[92] Caroline Arcanjo, Sandrine Frelon, Olivier Armant, Luc Camoin, Stéphane Audebert, Virginie Camilleri, Isabelle Cavalié, Christelle Adam-Guillermin, and Beatrice Gagnaire. Insights into the modes of action of tritium on the early-life stages of zebrafish, danio rerio, using transcriptomic and proteomic analyses. *Journal of Environmental Radioactivity*, 261:107141, 2023.

[93] Olivia Weeks, Bess M. Miller, Brian J. Pepe-Mooney, Isaac M. Oderberg, Scott H. Freeburg, Colton J. Smith, Trista E. North, and Wolfram Goessling. Embryonic alcohol exposure disrupts the ubiquitin-proteasome system. *JCI Insight*, 7(23), 2022.

[94] Cole D. English, Kira J. Kazi, Isaac Konig, Emma Ivantsova, Christopher L. Souders Ii, and Christopher J. Martyniuk. In silico microrna network data in zebrafish after antineoplastic ifosfamide exposure. *Data in Brief*, 48:109099, 2023.

[95] Sarah J. Patuel, Cole English, Victoria Lopez-Scarim, Isaac Konig, 2nd Souders, Christopher L., Emma Ivantsova, and Christopher J. Martyniuk. Dataset for diseases associated with exposure to broflanilide, a novel pesticide, in larval zebrafish (danio rerio). *Data Brief*, 50:109534, 2023.

[96] Simon Andrews. Fastqc: A quality control tool for high throughput sequence data, 2010.

[97] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, 17(1):3, 2011.

[98] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2012.

[99] Yang Liao, Gordon K. Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2013.

[100] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[101] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020.

[102] Aric Hagberg, Pieter J. Swart, and Daniel A. Schult. Exploring network structure, dynamics, and function using networkx, 2008.

[103] Ted G. Lewis. *Network science: theory and practice*. John Wiley & Sons, 2009.

[104] Jason E. McDermott, Ronald C. Taylor, Hyunjin Yoon, and Fred Heffron. Bottlenecks and hubs in inferred networks are important for virulence in salmonella typhimurium. *Journal of Computational Biology*, 16(2):169–180, 2009.

[105] Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019.

[106] Sashikanta Prusty, Srikanta Patnaik, and Sujit Kumar Dash. Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, 4, 2022.

[107] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[108] Yasunobu Nohara, Koutarou Matsumoto, Hidehisa Soejima, and Naoki Nakashima. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 214:106584, 2022.

[109] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1):281, 2019.

[110] Ryuichi Fukuda, Ruben Marín-Juez, Hadil El-Sammak, Arica Beisaw, Radhan Ramadass, Carsten Kuenne, Stefan Guenther, Anne Konzer, Aditya M. Bhagwat, Johannes Graumann, and Didier Y. Stainier. Stimulation of glycolysis promotes cardiomyocyte proliferation after injury in adult zebrafish. *EMBO Rep*, 21(8):e49752, 2020.

[111] Nelson S. Yee, Kristin Lorent, and Michael Pack. Exocrine pancreas development in zebrafish. *Developmental Biology*, 284(1):84–101, 2005.

[112] Jeroen Paardekooper Overman and Jeroen den Hertog. Zebrafish as a model to study ptps during development. *Methods*, 65(2):247–253, 2014.

[113] Yung-Che Tseng, Ruo-Dong Chen, Jay-Ron Lee, Sian-Tai Liu, Shyh-Jye Lee, and Pung-Pung Hwang. Specific expression and regulation of glucose transporters in zebrafish ionocytes. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 297(2):R275–R290, 2009.

[114] Liping Zhang, Zihong Chen, Ying Wang, David J. Tweardy, and William E. Mitch. Stat3 activation induces insulin resistance via a muscle-specific e3 ubiquitin ligase fbxo40. *American Journal of Physiology-Endocrinology and Metabolism*, 318(5):E625–E635, 2020.

[115] Byungyoung Ahn, Shibio Wan, Natasha Jaiswal, Rick B. Vega, Donald E. Ayer, Paul M. Titchenell, Xianlin Han, Kyoung J. Won, and Daniel P. Kelly. Mondoa drives muscle lipid accumulation and insulin resistance. *JCI Insight*, 5(15), 2019.

[116] Christiane Klec, Gabriela Ziomek, Martin Pichler, Roland Malli, and Wolfgang F. Graier. Calcium signaling in ß-cell physiology and pathology: A revisit. *International Journal of Molecular Sciences*, 20(24):6110, 2019.

[117] Pia Zilleßen, Jennifer Celner, Anita Kretschmann, Alexander Pfeifer, Kurt Racké, and Peter Mayer. Metabolic role of dipeptidyl peptidase 4 (dpp4) in primary human (pre)adipocytes. *Scientific Reports*, 6(1):23074, 2016.

[118] Richard Kin Ting Kam, Yi Deng, Yonglong Chen, and Hui Zhao. Retinoic acid synthesis and functions in early embryonic development. *Cell & Bioscience*, 2(1):11, 2012.

[119] Edward J. Lammer, Diane T. Chen, Richard M. Hoar, Narsingh D. Agnish, Paul J. Benke, John T. Braun, Cynthia J. Curry, Paul M. Fernhoff, Art W. Grix, Ira T. Lott, James M. Richard, and Shyan C. Sun. Retinoic acid embryopathy. *New England Journal of Medicine*, 313(14):837–841, 1985.

[120] Henry M. Sucov and Ronald M. Evans. Retinoic acid and retinoic acid receptors in development. *Molecular Neurobiology*, 10(2):169–184, 1995.

[121] Pedro J. Gomez-Pinilla, Pedro J. Camello, and María J. Pozo. Pancreatic calcium signaling: Role in health and disease. *Pancreatology*, 9(4):329–333, 2009.

[122] Shimpei Fujimoto, Koichiro Nabe, Mihoko Takehiro, Makiko Shimodahira, Mariko Kajikawa, Tomomi Takeda, Eri Mukai, Nobuya Inagaki, and Yutaka Seino. Impaired metabolism–secretion coupling in pancreatic β-cells: Role of determinants of mitochondrial atp production. *Diabetes Research and Clinical Practice*, 77(3, Supplement):S2–S10, 2007.

[123] Marjo J. Den Broeder, Victoria A. Kopylova, Leonie M. Kamminga, and Juliette Legler. Zebrafish as a model to study the role of peroxisome proliferating-activated receptors in adipogenesis and obesity. *PPAR Research*, 2015:358029, 2015.

[124] Olivia Venezia, Sadia Islam, Christine Cho, Alicia R. Timme-Laragy, and Karilyn E. Sant. Modulation of ppar signaling disrupts pancreas development in the zebrafish, danio rerio. *Toxicology and Applied Pharmacology*, 426:115653, 2021.

[125] Julia B. Ward, Sarah S. Casagrande, and Catherine C. Cowie. Urinary phenols and parabens and diabetes among us adults, nhanes 2005-2014. *Nutr Metab Cardiovasc Dis*, 30(5):768–776, 2020.

[126] M. Bortolini, M. B. Wright, M. Bopst, and B. Balas. Examining the safety of ppar agonists - current trends and future prospects. *Expert Opin Drug Saf*, 12(1):65–79, 2013.

[127] Rebecca Whitehead, Haiyan Guan, Edith Arany, Maria Cernea, and Kaiping Yang. Prenatal exposure to bisphenol a alters mouse fetal pancreatic morphology and islet composition. *Hormone Molecular Biology and Clinical Investigation*, 25(3):171–179, 2016.

[128] Akira Horii, Shuichi Nakatsuru, Yasuo Miyoshi, Shigetoshi Ichii, Hiroki Nagase, Hiroshi Ando, Akio Yanagisawa, Eiju Tsuchiya, Yo Kato, and Yusuke Nakamura. Frequent somatic mutations of the apc gene in human pancreatic cancer1. *Cancer Research*, 52(23):6696–6698, 1992.

[129] Qiansheng Huang and Qionghua Chen. Mediating roles of ppars in the effects of environmental chemicals on sex steroids. *PPAR Research*, 2017(1):3203161, 2017.

[130] Xiaowei Niu, Jingjing Zhang, Lanlan Zhang, Yangfan Hou, Shuangshuang Pu, Aiai Chu, Ming Bai, and Zheng Zhang. Weighted gene co-expression network analysis identifies critical genes in the development of heart failure after acute myocardial infarction. *Frontiers in Genetics*, 10, 2019.

[131] Li Zhu, Ying Ding, Cho-Yi Chen, Lin Wang, Zhiguang Huo, SungHwan Kim, Christos Sotiriou, Steffi Oesterreich, and George C Tseng. Metadcn: meta-analysis framework for differential co-expression network detection with an application in breast cancer. *Bioinformatics*, 33(8):1121–1129, 2016.

[132] B Chen, P Greenside, H Paik, M Sirota, D Hadley, and AJ Butte. Relating chemical structure to cellular response: An integrative analysis of gene expression, bioactivity, and structural data across 11,000 compounds. *CPT: Pharmacometrics & Systems Pharmacology*, 4(10):576–584, 2015.

[133] Sebastian Schmeisser, Andrea Miccoli, Martin von Bergen, Elisabet Berggren, Albert Braeuning, Wibke Busch, Christian Desaintes, Anne Gourmelon, Roland Grafström, Joshua Harrill, Thomas Hartung, Matthias Herzler, George E. N. Kass, Nicole Kleinstreuer, Marcel Leist, Mirjam Luijten, Philip Marx-Stoelting, Oliver Poetz, Bennard van Ravenzwaay, Rob Roggeband, Vera Rogiers, Adrian Roth, Pascal Sanders, Russell S. Thomas, Anne Marie Vinggaard, Mathieu Vinken, Bob van de Water, Andreas Luch, and Tewes Tralau. New approach methodologies in human regulatory toxicology – not if, but how and when! *Environment International*, 178:108082, 2023.

[134] Piper R. Hunt. The c. elegans model in toxicity testing. *J Appl Toxicol*, 37(1):50–59, 2017.

[135] Matthew D. Rand. Drosophotoxicology: The growing potential for drosophila in neurotoxicology. *Neurotoxicology and Teratology*, 32(1):74–83, 2010.

[136] Aya Takesono, Tetsuhiro Kudoh, and Charles R. Tyler. Application of transgenic zebrafish models for studying the effects of estrogenic endocrine disrupting chemicals on embryonic brain development. *Front Pharmacol*, 13:718072, 2022.

[137] Arielle Beaulieu. *California Drinking Water Contaminants and Adverse Birth Outcomes*. Thesis, San Diego State University, 2023.

[138] Daniel M. Ely and Anne K. Driscoll. Infant mortality in the united states, 2022: Data from the period linked birth/infant death file. *Natl Vital Stat Rep*, (5), 2024.

[139] Russell S. Kirby. The prevalence of selected major birth defects in the united states. *Seminars in Perinatology*, 41(6):338–344, 2017.

[140] Gary M. Shaw and Richard H. Finnel. *How do Gene-Environment Interactions Affect the Risk of Birth Defects?* Society for Birth Defects Research and Preventio, 2018.

[141] J. Yang, S. L. Carmichael, M. Canfield, J. Song, G. M. Shaw, and the National Birth Defects Prevention Study. Socioeconomic status in relation to selected birth defects in a large multicentered us case-control study. *American Journal of Epidemiology*, 167(2):145–154, 2007.

[142] Kirsten S. Almberg, Mary E. Turyk, Rachael M. Jones, Kristin Rankin, Sally Freels, Judith M. Graber, and Leslie T. Stayner. Arsenic in drinking water and adverse birth outcomes in ohio. *Environmental Research*, 157:52–59, 2017.

[143] Sarah C. Tinker, Suzanne Gilboa, Jennita Reefhuis, Mary M. Jenkins, Marcy Schaeffer, and Cynthia A. Moore. Challenges in studying modifiable risk factors for birth defects. *Current Epidemiology Reports*, 2(1):23–30, 2015.

[144] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[145] Marko Gosak, Rene Markovič, Jurij Dolenšek, Marjan Slak Rupnik, Marko Marhl, Andraž Stožer, and Matjaž Perc. Network science of biological systems at different scales: A review. *Physics of Life Reviews*, 24:118–135, 2018.

[146] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D. Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, Lingyun Yang, Cheryl Wolting, Ian Donaldson, Søren Schandorff, Juanita Shewnarane, Mai Vo, Joanne Taggart, Marilyn Goudreault, Brenda Muskat, Cris Alfarano, Danielle Dewar, Zhen Lin, Katerina Michalickova, Andrew R. Willems, Holly Sassi, Peter A. Nielsen, Karina J. Rasmussen, Jens R. Andersen, Lene E. Johansen, Lykke H. Hansen, Hans Jespersen, Alexandre Podtelejnikov, Eva Nielsen, Janne Crawford, Vibeke Poulsen, Birgitte D. Sørensen, Jesper Matthiesen, Ronald C. Hendrickson, Frank Gleeson, Tony Pawson, Michael F. Moran, Daniel Durocher, Matthias Mann, Christopher W. V. Hogue, Daniel Figeys, and Mike Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[147] Pei Wang. Network biology: Recent advances and challenges. *GPD*, 1(2), 2022.

[148] Alberto Aleta and Yamir Moreno. Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics*, 10(1):45–62, 2019.

[149] Oriol Artime, Barbara Benigni, Giulia Bertagnolli, Valeria d'Andrea, Riccardo Gallotti, Arsham Ghavasieh, Sebastian Raimondo, and Manlio De Domenico. Multilayer network science: From cells to societies, 2022.

[150] Nitzan Shalom Artzi, Smadar Shilo, Eran Hadar, Hagai Rossman, Shiri Barbash-Hazan, Avi Ben-Haroush, Ran D. Balicer, Becca Feldman, Arnon Wiznitzer, and Eran Segal. Prediction of gestational diabetes based on nationwide electronic health records. *Nature Medicine*, 26(1):71–76, 2020.

[151] Jean-Baptiste Guimbaud, Alexandros P. Siskos, Amrit Kaur Sakhi, Barbara Heude, Eduard Sabidó, Eva Borràs, Hector Keun, John Wright, Jordi Julvez, Jose Urquiza, Kristine Bjerve Gützkow, Leda Chatzi, Maribel Casas, Mariona Bustamante, Mark

Nieuwenhuijsen, Martine Vrijheid, Mónica López-Vicente, Montserrat de Castro Pascual, Nikos Stratakis, Oliver Robinson, Regina Grazuleviciene, Remy Slama, Silvia Alemany, Xavier Basagaña, Marc Plantevit, Rémy Cazabet, and Léa Maitre. Machine learning-based health environmental-clinical risk scores in european children. *Communications Medicine*, 4(1):98, 2024.

[152] Zhoupeng Ren, Jun Zhu, Yanfang Gao, Qian Yin, Maogui Hu, Li Dai, Changfei Deng, Lin Yi, Kui Deng, Yanping Wang, Xiaohong Li, and Jinfeng Wang. Maternal exposure to ambient pm10 during pregnancy increases the risk of congenital heart defects: Evidence from machine learning models. *Science of The Total Environment*, 630:1–10, 2018.

[153] Xiaoyi Raymond Gao, Marion Chiariglione, Ke Qin, Karen Nuytemans, Douglas W. Scharre, Yi-Ju Li, and Eden R. Martin. Explainable machine learning aggregates polygenic risk scores and electronic health records for alzheimer's disease prediction. *Scientific Reports*, 13(1):450, 2023.

[154] Fuliang Yi, Hui Yang, Durong Chen, Yao Qin, Hongjuan Han, Jing Cui, Wenlin Bai, Yifei Ma, Rong Zhang, and Hongmei Yu. Xgboost-shap-based interpretable diagnostic framework for alzheimer's disease. *BMC Medical Informatics and Decision Making*, 23(1):137, 2023.

[155] California Office of Environmental Health Hazard Assessment (OEHHA). Calenviroscreen 4.0, 2023.

[156] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 2nd edition edition, 1988.

[157] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[158] Barret A. Monchka, Carson K. Leung, Nathan C. Nickel, and Lisa M. Lix. The effect of disease co-occurrence measurement on multimorbidity networks: a population-based study. *BMC Medical Research Methodology*, 22(1):165, 2022.

[159] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[160] Steve Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer New York Dordrecht Heidelberg London, 2011.

[161] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system, 2016.

[162] Sergiu Hart. *Shapley Value*, pages 210–216. Palgrave Macmillan UK, London, 1989.

[163] Keita Ebisu, Jesse D. Berman, and Michelle L. Bell. Exposure to coarse particulate matter during gestation and birth weight in the u.s. *Environment International*, 94:519–524, 2016.

[164] Zhiqiang Nie, Yanji Qu, Fengzhen Han, Erin M. Bell, Jian Zhuang, Jimei Chen, Melissa François, Emily Lipton, Rosemary Matale, Weilun Cui, Qianhong Liang, Xiangzhang Lu, Huiwen Huang, Junfeng Lv, Yanqiu Ou, Jinzhuang Mai, Yong Wu, Xiangmin Gao, Yating Huang, Shao Lin, and Xiaoqing Liu. Evaluation of interactive effects between paternal alcohol consumption and paternal socioeconomic status and environmental exposures on congenital heart defects. *Birth Defects Research*, 112(16):1273–1286, 2020.

[165] C R Wasserman, G M Shaw, S Selvin, J B Gould, and S L Syme. Socioeconomic status, neighborhood social conditions, and neural tube defects. *American Journal of Public Health*, 88(11):1674–1680, 1998.

[166] Derek A. Chapman, Caroline C. Stampfel, Joann N. Bodurtha, Kelley M. Dodson, Arti Pandya, Kathleen B. Lynch, and Russell S. Kirby. Impact of co-occurring birth defects on the timing of newborn hearing screening and diagnosis. *American Journal of Audiology*, 20(2):132–139, 2011.

[167] Anna M. H. Korver, Richard J. H. Smith, Guy Van Camp, Mark R. Schleiss, Maria A. K. Bitner-Glindzicz, Lawrence R. Lustig, Shin-ichi Usami, and An N. Boudewyns. Congenital hearing loss. *Nature Reviews Disease Primers*, 3(1):16094, 2017.

[168] Paul M. Lantos, Gabriela Maradiaga-Panayotti, Xavier Barber, Eileen Raynor, Debara Tucci, Kate Hoffman, Sallie R. Permar, Pearce Jackson, Brenna L. Hughes, Amy Kind, and Geeta K. Swamy. Geographic and racial disparities in infant hearing loss. *Otolaryngology–Head and Neck Surgery*, 159(6):1051–1057, 2018.

[169] Blake Smith, Jessica Zhang, Gina Nhu Pham, Keerthana Pakanati, Nikhila Raol, Julina Ongkasuwan, and Samantha Anne. Effects of socioeconomic status on children with hearing loss. *International Journal of Pediatric Otorhinolaryngology*, 116:114–117, 2019.

[170] Mathew Christensen, Vickie Thomson, and G. William Letson. Evaluating the reach of universal newborn hearing screening in colorado. *American Journal of Preventive Medicine*, 35(6):594–597, 2008.

[171] William Cava, Christopher Bauer, Jason H. Moore, and Sarah A. Pendergrass. Interpretation of machine learning predictions for patient outcomes in electronic health records. *AMIA Annu Symp Proc*, 2019:572–581, 2019.

[172] Abdullah Alanazi. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30:100924, 2022.

[173] Bárbara Avelar-Pereira, Michael E. Belloy, Ruth O'Hara, S. M. Hadi Hosseini, and Initiative for the Alzheimer's Disease Neuroimaging. Decoding the heterogeneity of

alzheimer's disease diagnosis and progression using multilayer networks. *Molecular Psychiatry*, 28(6):2423–2432, 2023.

[174] Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, 2014.

[175] Aylin Alin. Multicollinearity. *WIREs Computational Statistics*, 2(3):370–374, 2010.

[176] Margaret A. Honein, Russell S. Kirby, Robert E. Meyer, Jian Xing, Nyasha I. Skerrette, Nataliya Yuskiv, Lisa Marengo, Joann R. Petrini, Michael J. Davidoff, Cara T. Mai, Charlotte M. Druschel, Samara Viner-Brown, Lowell E. Sever, and Network for the National Birth Defects Prevention. The association between major birth defects and preterm birth. *Maternal and Child Health Journal*, 13(2):164–175, 2009.

[177] Liang Yu, Dandan Zhou, Lin Gao, and Yunhong Zha. Prediction of drug response in multilayer networks based on fusion of multiomics data. *Methods*, 192:85–92, 2021.

[178] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning, 2020.

[179] California Office of Environmental Health Hazard Assessment (OEHHA). Sb 535 disadvantaged communities, n.d.

[180] California Office of Environmental Health Hazard Assessment (OEHHA). Calenviroscreen 4.0: Report on update to the california communities environmental health screening tool., 2021.

[181] Ann Aschengrau, Janice M. Weinberg, Patricia A. Janulewicz, Lisa G. Gallagher, Michael R. Winter, Veronica M. Vieira, Thomas F. Webster, and David M. Ozonoff. Prenatal exposure to tetrachloroethylene-contaminated drinking water and the risk of congenital anomalies: a retrospective cohort study. *Environmental Health*, 8(1):44, 2009.

[182] Lisa M. McKenzie, William Allshouse, and Stephen Daniels. Congenital heart defects and intensity of oil and gas well site activities in early pregnancy. *Environ Int*, 132:104949, 2019.

[183] Khaiwal Ravindra, Neha Chanana, and Suman Mor. Exposure to air pollutants and risk of congenital anomalies: A systematic review and metaanalysis. *Science of The Total Environment*, 765:142772, 2021.

[184] Bin Zhang, Shengwen Liang, Jinzhu Zhao, Zhengmin Qian, Bryan A. Bassig, Rong Yang, Yiming Zhang, Ke Hu, Shunqing Xu, Tongzhang Zheng, and Shaoping Yang. Maternal exposure to air pollutant pm2.5 and pm10 during pregnancy and risk of congenital heart defects. *J Expo Sci Environ Epidemiol*, 26(4):422–7, 2016.

[185] Yanji Qu, Xinlei Deng, Shao Lin, Fengzhen Han, Howard H. Chang, Yanqiu Ou, Zhiqiang Nie, Jinzhuang Mai, Ximeng Wang, Xiangmin Gao, Yong Wu, Jimei Chen, Jian Zhuang, Ian Ryan, and Xiaoqing Liu. Using innovative machine learning methods to screen and identify predictors of congenital heart diseases. *Frontiers in Cardiovascular Medicine*, 8, 2022.

[186] Benjamin A. Goldstein, Ann Marie Navar, and Rickey E. Carter. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European Heart Journal*, 38(23):1805–1814, 2016.

[187] Chris Barnett, James Christiansen, Monica Mills, Jordyn Lord, and Jared Parrish. Measuring misclassification and sample bias in passive surveillance systems: Improving prevalence estimates of critical congenital heart defects in state-based passive surveillance systems. *Birth Defects Research*, 116(8):e2386, 2024.

[188] Jessica C. Young, Mitchell M. Conover, and Michele Jonsson Funk. Measurement error and misclassification in electronic medical records: Methods to mitigate bias. *Current Epidemiology Reports*, 5(4):343–356, 2018.

[189] Benjamin Q. Huynh, Elizabeth T. Chin, Allison Koenecke, Derek Ouyang, Daniel E. Ho, Mathew V. Kiang, and David H. Rehkopf. Mitigating allocative tradeoffs and harms in an environmental justice data tool. *Nature Machine Intelligence*, 6(2):187–194, 2024.

[190] Chathurangi Shyalika, Ruwan Wickramarachchi, and Amit P. Sheth. A comprehensive survey on rare event prediction. *ACM Comput. Surv.*, 57(3):Article 70, 2024.

[191] Sheng Zhong, Jane Zhang, Jenny Jiao, Hongjian Zhu, Yunzhao Xing, and Li Wang. A machine learning case study to predict rare clinical event of interest: imbalanced data, interpretability, and practical considerations. *Journal of Biopharmaceutical Statistics*, pages 1–14, 2024.

Appendix A

Supplementary Material for Chapter 2

A.1 Supplementary Figures

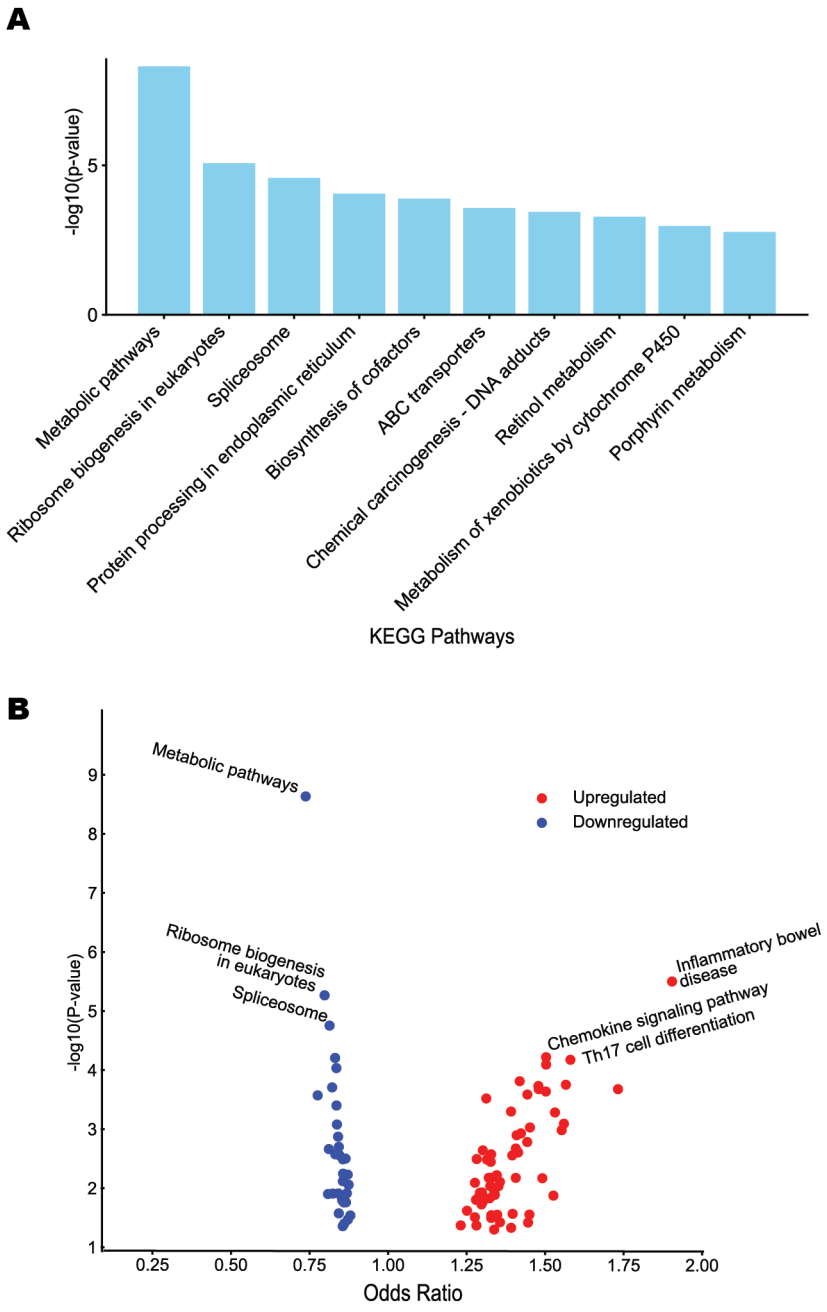

Figure A.1: Enrichplots module additional plotting options for the data presented in Figure 2.1 (A) Bar chart displaying the top 10 downregulated pathways. (B) Volcano plot displaying the relationship between the odds ratio and the p-value for upregulated and downregulated pathways.

Appendix B

Supplementary Material for Chapter 3

B.1 Supplementary Tables

Table B.1: Chemical properties for each chemical exposure curated from the Gene Expression Omnibus. Abbreviations include: Halogenated Organic Compounds (HOC), Perfluoroalkyl Substances (PFAS), Pesticide/Herbicide (P/H), Paraben (PAR). Pharmaceuticals/Drugs (P/D), Endocrine Disrupting Chemicals (EDC), and Solvents/Other (S/O).

Chemical Label	**Chemical Full Name**	**Chemical Short Name**	**CAS #**	**DTX ID**	**Chemical Type**	**GEO Accession**	**DOI**	**Age (dpf)**	**Micromolar Dose**	**Fish Strain**	**Exposure Time (hrs)**	**Repeated Exposure**
PCB11_20.2uM	3,3'-Dichlorobiphenyl	PCB-11	2050-67-1	DTXSID 70872817	HOC	GSE 118955	10.1016/ j.envpol. 2019.113027	4	20.1703	AB-AB_ Background	72	0
TCPM_50nM	Tris(4-chlorophenyl) methane	TCPM	27575-78-6	DTXSID 70181976	HOC	GSE 198106	10.1002/ bdr2.2132	4	0.05	AB-AB_ Background	72	1
TCPM_50nM _acute	Tris(4-chlorophenyl) methane	TCPM	27575-78-6	DTXSID 70181976	HOC	GSE 165920	10.1016/ j.aquatox. 2021.105815	4	0.05	AB-AB_ Background	72	1
TCPMOH_50nM	Tris(4-chlorophenyl) methanol	TCP MOH	3010-80-8	DTXSID 50184187	HOC	GSE 198106	10.1002/ bdr2.2132	4	0.05	AB-AB_ Background	72	1
TCPMOH_50nM _acute	Tris(4-chlorophenyl) methanol	TCP MOH	3010-80-8	DTXSID 50184187	HOC	GSE 165920	10.1016/ j.aquatox. 2021.105815	4	0.05	AB-AB_ Background	72	1
OBS_32uM	Sodium p-perfluorous nonenoxybenzene sulfonate	OBS	70829–87-7	DTXSID 601020833	PFAS	GSE 164074	10.1016/ j.scitotenv. 2021.145443	4	31.938	AB-TL	90	1
OBS_48uM	Sodium p-perfluorous nonenoxybenzene sulfonate	OBS	70829–87-7	DTXSID 601020833	PFAS	GSE 164074	10.1016/ j.scitotenv. 2021.145443	4	47.90649	AB-TL	90	1

Continued on next page

Table B.1 – continued from previous page

Chemical Label	Chemical Full Name	Chemical Short Name	CAS #	DTX ID	Chemical Type	GEO Accession	DOI	Age (dpf)	Micromolar Dose	Fish Strain	Exposure Time (hrs)	Repeated Exposure
PFBS_16uM	Perfluorobutanesulfonic Acid	PFBS	375-73-5	DTXSID 5030030	PFAS	GSE 114356	10.1093/toxsci/kfy237	4	16	AB-AB_Background	167	1
PFBS_32uM	Perfluorobutanesulfonic Acid	PFBS	375-73-5	DTXSID 5030030	PFAS	GSE 114356	10.1093/toxsci/kfy237	4	32	AB-AB_Background	167	1
PFOS_0.6uM_A	Perfluorooctanesulfonic acid	PFOS	1763-23-1	DTXSID 3031864	PFAS	GSE 144525	10.1016/j.envres.2020.109702	4	0.59984	AB-AB_Background	90	0
PFOS_4.1uM	Perfluorooctanesulfonic acid	PFOS	1763-23-1	DTXSID 3031864	PFAS	GSE 144525	10.1016/j.envres.2020.109702	4	4.11893	AB-AB_Background	90	0
PFOS_40uM	Perfluorooctanesulfonic acid	PFOS	1763-23-1	DTXSID 3031864	PFAS	GSE 164074	10.1016/j.scitotenv.2021.145443	4	39.9896	AB-TL	90	1
Boscalid_8.7nM	Boscalid	Boscalid	188425-85-6	DTXSID 6034392	P/H	GSE 232317		4	0.00874	AB-AB_Background	92	1
M510F01_8.4nM	M510F01	M510F01	661463-87-2	DTXSID 201044409	P/H	GSE 232317		4	0.0084	AB-AB_Background	92	1
CBD_0.5uM	Cannabidiol	CBD	13956-29-1	DTXSID 00871959	P/D	GSE 164128	10.1093/toxsci/kfab046	4	0.5	WT-EKW Background	94	0
THC_4uM	Tetrahydrocannabinol	THC	1972-08-3	DTXSID 6021327	P/D	GSE 164128	10.1093/toxsci/kfab046	4	4	WT-EKW Background	94	0

Continued on next page

Table B.1 – continued from previous page

Chemical Label	Chemical Full Name	Chemical Short Name	CAS #	DTX ID	Chemical Type	GEO Accession	DOI	Age (dpf)	Micromolar Dose	Fish Strain	Exposure Time (hrs)	Repeated Exposure
BPA_0.4uM	Bisphenol A	BPA	80-05-7	DTXSID 7020182	EDC	GSE 113676	10.1016/ j.envpol. 2018.09.043	5	0.438	WT	72	1
BPA_4.4uM	Bisphenol A	BPA	80-05-7	DTXSID 7020182	EDC	GSE 113676	10.1016/ j.envpol. 2018.09.043	5	4.38	WT	72	1
BPA_17.5uM	Bisphenol A	BPA	80-05-7	DTXSID 7020182	EDC	GSE 113676	10.1016/ j.envpol. 2018.09.043	5	17.52	WT	72	1
TBT_0.1uM	Tributyltin	TBT	1461-22-9	DTXSID 3027403	EDC	GSE 139599	10.1016/ j.jhazmat. 2020.122881	5	0.1	AB-AB_ Background	72	1
TBT_30nM	Tributyltin	TBT	1461-22-9	DTXSID 3027403	EDC	GSE 139599	10.1016/ j.jhazmat. 2020.122881	5	0.03	AB-AB_ Background	72	1
TBT_3nM	Tributyltin	TBT	1461-22-9	DTXSID 3027403	EDC	GSE 139599	10.1016/ j.jhazmat. 2020.122881	5	0.003	AB-AB_ Background	72	1
PCB126_0.3nM	3,3',4,4',5-Pentachlorobiphenyl	PCB-126	57465-28-8	DTXSID 3032179	HOC	GSE 98741	10.1093/ toxsci/ kfx192	5	0.0003	TL	20	0
PCB126_1.2nM	3,3',4,4',5-Pentachlorobiphenyl	PCB-126	57465-28-8	DTXSID 3032179	HOC	GSE 98741	10.1093/ toxsci/ kfx192	5	0.0012	TL	20	0
PCB153_0.1uM	2,2',4,4',5,5'-hexachlorobiphenyl	PCB-153	35065-27-1	DTXSID 2032180	HOC	GSE 93310	10.1093/ toxsci/ kfz217	5	0.1	TL	116	1
PCB153_10uM	2,2',4,4',5,5'-hexachlorobiphenyl	PCB-153	35065-27-1	DTXSID 2032180	HOC	GSE 93310	10.1093/ toxsci/ kfz217	5	10	TL	116	1

Continued on next page

Table B.1 – continued from previous page

Chemical Label	Chemical Full Name	Chemical Short Name	CAS #	DTX ID	Chemical Type	GEO Accession	DOI	Age (dpf)	Micromolar Dose	Fish Strain	Exposure Time (hrs)	Repeated Exposure
PCB153_1uM	2,2',4,4',5,5'-hexachlorobiphenyl	PCB-153	35065-27-1	DTXSID 2032180	HOC	GSE 93310	10.1093/toxsci/kfz217	5	1	TL	116	1
EtP_1.5uM	Ethyl-Parahydroxybenzoates	EtP	120-47-8	DTXSID 9022528	PAR	GSE 239773	10.1016/j.ecoenv.2023.115704	5	1.452	AB-AB_Background	120	0
EtP_7.3uM	Ethyl-Parahydroxybenzoates	EtP	120-47-8	DTXSID 9022528	PAR	GSE 239773	10.1016/j.ecoenv.2023.115704	5	7.26	AB-AB_Background	120	0
EtP_72.6uM	Ethyl-Parahydroxybenzoates	EtP	120-47-8	DTXSID 9022528	PAR	GSE 239773	10.1016/j.ecoenv.2023.115704	5	72.6	AB-AB_Background	120	0
MtP_15.4uM	Methyl-Parahydroxybenzoates	MtP	99-76-3	DTXSID 4022529	PAR	GSE 239773	10.1016/j.ecoenv.2023.115704	5	15.41	AB-AB_Background	120	0
MtP_154.1uM	Methyl-Parahydroxybenzoates	MtP	99-76-3	DTXSID 4022529	PAR	GSE2 39773	10.1016/j.ecoenv.2023.115704	5	154.1	AB-AB_Background	120	0
MtP_3.1uM	Methyl-Parahydroxybenzoates	MtP	99-76-3	DTXSID 4022529	PAR	GSE 239773	10.1016/j.ecoenv.2023.115704	5	3.082	AB-AB_Background	120	0
PrP_0.5uM	Propyl- Parahydroxybenzoates	PrP	94-13-3	DTXSID 4022527	PAR	GSE 239773	10.1016/j.ecoenv.2023.115704	5	0.484	AB-AB_Background	120	0
PrP_2.4uM	Propyl- Parahydroxybenzoates	PrP	94-13-3	DTXSID 4022527	PAR	GSE 239773	10.1016/j.ecoenv.2023.115704	5	2.42	AB-AB_Background	120	0
PrP_24.2uM	Propyl- Parahydroxybenzoates	PrP	94-13-3	DTXSID 4022527	PAR	GSE 239773	10.1016/j.ecoenv.2023.115704	5	24.2	AB-AB_Background	120	0

Continued on next page

Table B.1 – continued from previous page

Chemical Label	Chemical Full Name	Chemical Short Name	CAS #	DTX ID	Chemical Type	GEO Accession	DOI	Age (dpf)	Micromolar Dose	Fish Strain	Exposure Time (hrs)	Repeated Exposure
PFOS_60nM	Perfluorooctanesulfonic acid	PFOS	1763-23-1	DTXSID3031864	PFAS	GSE125072	10.1016/j.scitotenv.2019.04.200	5	0.05998	AB-AB_Background	72	1
PFOS_0.6uM_B	Perfluorooctanesulfonic acid	PFOS	1763-23-1	DTXSID3031864	PFAS	GSE125072	10.1016/j.scitotenv.2019.04.200	5	0.5998	AB-AB_Background	72	1
PFOS_2uM	Perfluorooctanesulfonic acid	PFOS	1763-23-1	DTXSID3031864	PFAS	GSE125072	10.1016/j.scitotenv.2019.04.200	5	1.999	AB-AB_Background	72	1
Cortisol_1uM	Cortisol	Cortisol	50-23-7	DTXSID7020714	P/D	GSE80221	10.1242/bio.020065	5	1	AB-AB_Background	120	1
Triptolide_1.6uM	Triptolide	Triptolide	38748-32-2	DTXSID5041144	P/D	GSE146199	10.1093/toxsci/kfx144	5	1.6	WIK	6	0
Afid_1.7nM	Afidopyropen	Afid	915972-17-7	DTXSID00896889	P/H	GSE224522		7	0.001684	AB-TL	164	1
Broflanilide_15.1nM	Broflanilide	broflanilide	1207727-04-5	DTXSID50894815	P/H	GSE236484	10.1016/j.dib.2023.109534	7	0.0151	AB-TL	164	1
Broflanilide_1.5nM	Broflanilide	broflanilide	1207727-04-5	DTXSID50894815	P/H	GSE236484	10.1016/j.dib.2023.109534	7	0.00151	AB-TL	164	1
IFO_0.4uM	Ifosfamide	IFO	3778-73-2	DTXSID7020760	P/H	GSE217875	10.1016/j.dib.2023.109099	7	0.383	AB-TL	164	1
IFO_3.8nM	Ifosfamide	IFO	3778-73-2	DTXSID7020760	P/H	GSE217875	10.1016/j.dib.2023.109099	7	0.00383	AB-TL	164	1

Continued on next page

Table B.1 – continued from previous page

Chemical Label	Chemical Full Name	Chemical Short Name	CAS #	DTX ID	Chemical Type	GEO Accession	DOI	Age (dpf)	Micromolar Dose	Fish Strain	Exposure Time (hrs)	Repeated Exposure
AVS_0.2nM	Atorvastatin	AVS	134523-00-5	DTXSID8029868	P/D	GSE224522		7	0.000179	AB-TL	164	1
AVS_1.8nM	Atorvastatin	AVS	134523-00-5	DTXSID8029868	P/D	GSE224522		7	0.00179	AB-TL	164	1
CIT_3.1uM	Citalopram	CIT	59729-33-8	DTXSID8022826	P/D	GSE217875	10.1016/j.dib.2023.109099	7	3.0826	AB-TL	164	1
CIT_30.8nM	Citalopram	CIT	59729-33-8	DTXSID8022826	P/D	GSE217875	10.1016/j.dib.2023.109099	7	0.0308	AB-TL	164	1
EtOH_217mM	Ethanol	EtOH	64-17-5	DTXSID9020584	S/O	GSE172111	10.1172/jci.insight.156914	7	217000	AB-AB_Background	108	0
HTO_3.3uM	Tritiated Water	HTO	14940-65-9	DTXSID80894747	S/O	GSE168194	10.1016/j.jenvrad.2023.107141	7	3.3	AB-AB_Background	165	1

Table B.2: Chemical properties for each chemical exposure curated from the Gene Expression Omnibus curated from the Environmental Protection Agency Computational Toxicology Database. Column title abbreviations and associated units include: Average Mass (g/mol) (AM), Monoisotpic Mass (g/mol) (MM), Atmos. Hydroxylation Rate (cm3/molecule*sec) (AHR), Bioconcentration Factor (L/kg) (BF), Biodeg. Half-Life (days) (BHL), Boiling Point (°C) (BP), Density (g/cmˆ3) (D), Fish Biotrans. Half-Life (Km) (days) (FHL), Flash Point (°C) (FP), Henry's Law (atm-m3/mole) (HL), Index of Refraction (IOR), LogKoa: Octanol-Air (Log Koa), LogKow: Octanol-Water (Log Kow), Melting Point (°C) (MP), Molar Refractivity (cmˆ3) (MR), Molar Volume (cmˆ3) (MV), Polarizability (Åˆ3) (P), Soil Adsorp. Coeff. (L/kg) (Koc), Surface Tension (dyn/cm) (ST), Vapor Pressure (mmHg) (VP), Water Solubility (mol/L) (WS)

Chemical Label	**AM**	**MM**	**AHR**	**BF**	**BHL**	**BP**	**D**	**FHL**	**FP**	**HL**	**IOR**	**Log D5.5**	**Log D7.4**	**Log Koa**	**Log Kow**	**MP**	**MR**	**MV**	**P**	**Koc**	**ST**	**VP**	**WS**
PCB11_20.2uM	223.1	222.0	4.1e-12	6460	18.6	319	1.25	28.2	143	2.30e-04	1.6	5.3	5.3	7.4	5.1	55	60.6	179	24	11200	42.8	6.4e+04	1.39e-06
TCPM_50nM	347.7	346.0	8.3e-12	5190	38	409	1.31	50.1	251	1.48e-06	1.6	6.8	6.8	10.5	7.3	139	94.7	269	37.5	45700	45.9	3.72e-07	3.71e-07
TCPM_50nM_acute	347.7	346.0	8.3e-12	5190	38	409	1.31	50.1	251	1.48e-06	1.6	6.8	6.8	10.5	7.3	139	94.7	269	37.5	45700	45.9	3.72e-07	3.71e-07
TCPMOH_50nM	363.7	362.0	1.7e-11	2390	63.1	442	1.38	3.9	245	2.57e-10	1.6	4.6	4.6	10.5	6.3	166	95.8	266	38	1950	50.3	2.7e-08	1.72e-06
TCPMOH_50nM_acute	347.7	346.0	1.7e-11	2390	63.1	442	1.38	3.9	245	2.57e-10	1.6	4.6	4.6	10.5	6.3	166	95.8	266	38	1950	50.3	2.7e-08	1.72e-06
OBS_32uM	626.2	625.9	8.7e-13	66.1	5.75	112	604	3.2	1	1 e-09	760	4.7	2.8	8.6	7.0	169	25	0.405	0.135	1.32e+05	0.5	8.11e-06	16.3
OBS_48uM	626.2	625.9	8.7e-13	66.1	5.75	112	604	3.2	1	1 e-09	760	4.7	2.8	8.6	7.0	169	25	0.405	0.135	1.32e+05	0.5	8.11e-06	16.3
PFBS_16uM	300.1	300.0	2.2e-15	6.82	4.47	211	1.83	0.1	190	2.95e-10	1.3	-4.0	-5.9	4.2	2.8	36.9	32	162	12.7	288	23.4	0.104	5.02e-03

Continued on next page

Table B.2 – continued from previous page

Chemical Label	AM	MM	AHR	BF	BHL	BP	D	FHL	FP	HL	IOR	Log D5.5	Log D7.4	Log Koa	Log Kow	MP	MR	MV	P	Koc	ST	VP	WS
PFBS_32uM	300.1	300.0	2.2e-15	6.82	4.47	211	1.83	0.1	190	2.95e-10	1.3	-4.0	-5.9	4.2	2.8	36.9	32	162	12.7	288	23.4	0.104	5.02e-03
PFOS_0.6uM_A	500.1	499.9	2 e-15	464	4.9	229	1.84	2.7	190	1.82e-11	1.3	-1.5	-3.4	4.8	6.0	51.9	51.5	272	20.4	1460	19.6	2.45e-06	4.09e-04
PFOS_4.1uM	500.1	499.9	2 e-15	464	4.9	229	1.84	2.7	190	1.82e-11	1.3	-1.5	-3.4	4.8	6.0	51.9	51.5	272	20.4	1460	19.6	2.45e-06	4.09e-04
PFOS_40uM	500.1	499.9	2 e-15	464	4.9	229	1.84	2.7	190	1.82e-11	1.3	-1.5	-3.4	4.8	6.0	51.9	51.5	272	20.4	1460	19.6	2.45e-06	4.09e-04
Boscalid_8.7nM	343.2	342.0	1.1e-11	95.3	6.03	447	1.38	1.4	242	2.69e-09	1.7	3.0	3.0	10.3	4.2	178	93.3	251	37	928	56.1	1.04e-08	6.56e-06
M510F01_8.4nM	359.2	358.0	1.1e-11	81.3	20	189	359	0.4	3	3.55e-09	760	3.3	3.4	9.6	3.9	189	25	0.75	0.167	794	0.1	1.41e-09	18.9
CBD_0.5uM	314.5	314.5	1.3e-10	912	36.3	424	1.02	1.9	206	2.34e-09	1.5	6.4	6.4	9.3	6.4	67	97	307	38.5	6030	39.6	1.78e-09	3.21e-05
THC_4uM	314.5	314.2	4.7e-11	334	85.1	382	1.04	3.0	165	8.32e-10	1.5	7.2	7.0	10.8	7.4	134	95.5	310	37.9	6.76e+04	37.4	6.01e-08	3.76e-06
BPA_0.4uM	228.3	228.1	1.6e-11	94.6	15.1	362	1.17	1.9	190	1.26e-07	1.6	3.3	3.3	8.4	3.5	132	68.2	200	27	1340	46.0	5.34e-07	7.49e-04
BPA_4.4uM	228.3	228.1	1.6e-11	94.6	15.1	362	1.17	1.9	190	1.26e-07	1.6	3.3	3.3	8.4	3.5	132	68.2	200	27	1340	46.0	5.34e-07	7.49e-04
BPA_17.5uM	228.3	228.1	1.6e-11	94.6	15.1	362	1.17	1.9	190	1.26e-07	1.6	3.3	3.3	8.4	3.5	132	68.2	200	27	1340	46.0	5.34e-07	7.49e-04

Continued on next page

Table B.2 – continued from previous page

Chemical Label	AM	MM	AHR	BF	BHL	BP	D	FHL	FP	HL	IOR	Log D5.5	Log D7.4	Log Koa	Log Kow	MP	MR	MV	P	Koc	ST	VP	WS
TBT_0.1uM	325.5	326.1	1.1e-11	493	5.75	282	1.32	1.9	129	2.95e-09	1.6	3.0	2.9	8.6	3.9	42.6	64.4	226	25.5	1460	42.6	3.83e-03	1.7e-06
TBT_30nM	325.5	326.1	1.1e-11	493	5.75	282	1.32	1.9	129	2.95e-09	1.6	3.0	2.9	8.6	3.9	42.6	64.4	226	25.5	1460	42.6	3.83e-03	1.7e-06
TBT_3nM	325.5	326.1	1.1e-11	493	5.75	282	1.32	1.9	129	2.95e-09	1.6	3.0	2.9	8.6	3.9	42.6	64.4	226	25.5	1460	42.6	3.83e-03	1.7e-06
PCB126_0.3nM	326.4	323.9	6.6e-13	51600	16.2	379	1.52	251	193	5.89e-05	1.6	6.6	6.6	10.2	6.5	113	75.3	214	29.9	157000	46.5	5.96e-06	2.63e-08
PCB126_1.2nM	326.4	323.9	6.6e-13	51600	16.2	379	1.52	251	193	5.89e-05	1.6	6.6	6.6	10.2	6.5	113	75.3	214	29.9	157000	46.5	5.96e-06	2.63e-08
PCB153_0.1uM	360.9	357.8	6.6e-13	184000	20	395	1.6	355	191	2.34e-05	1.6	6.6	6.6	9.7	6.8	127	80.2	226	31.8	3.89e+05	48.0	2.35e-06	7.25e-09
PCB153_10uM	360.9	357.8	6.6e-13	184000	20	395	1.6	355	191	2.34e-05	1.6	6.6	6.6	9.7	6.8	127	80.2	226	31.8	3.89e+05	48.0	2.35e-06	7.25e-09
PCB153_1uM	360.9	357.8	6.6e-13	184000	20	395	1.6	355	191	2.34e-05	1.6	6.6	6.6	9.7	6.8	127	80.2	226	31.8	3.89e+05	48.0	2.35e-06	7.25e-09
EtP_1.5uM	166.2	166.1	1.1e-11	5.37	3.55	283	1.16	0.1	115	2.95e-09	1.5	2.5	2.5	7.7	2.5	70.8	44.5	142	17.7	162	41.1	7.59e-04	9.89e-03
EtP_7.3uM	166.2	166.1	1.1e-11	5.37	3.55	283	1.16	0.1	115	2.95e-09	1.5	2.5	2.5	7.7	2.5	70.8	44.5	142	17.7	162	41.1	7.59e-04	9.89e-03
EtP_72.6uM	166.2	166.1	1.1e-11	5.37	3.55	283	1.16	0.1	115	2.95e-09	1.5	2.5	2.5	7.7	2.5	70.8	44.5	142	17.7	162	41.1	7.59e-04	9.89e-03

Continued on next page

Table B.2 – continued from previous page

Chemical Label	AM	MM	AHR	BF	BHL	BP	D	FHL	FP	HL	IOR	Log D5.5	Log D7.4	Log Koa	Log Kow	MP	MR	MV	P	Koc	ST	VP	WS
MtP_15.4uM	152.1	152.0	1.1e-11	5.11	3.55	260	1.2	0.1	115	2.95e-09	1.6	2.0	1.9	8.6	2.0	69.4	39.9	126	15.8	56.2	42.3	5.55e-03	2.75e-02
MtP_154.1uM	152.1	152.0	1.1e-11	5.11	3.55	260	1.2	0.1	115	2.95e-09	1.6	2.0	1.9	8.6	2.0	69.4	39.9	126	15.8	56.2	42.3	5.55e-03	2.75e-02
MtP_3.1uM	152.1	152.0	1.1e-11	5.11	3.55	260	1.2	0.1	115	2.95e-09	1.6	2.0	1.9	8.6	2.0	69.4	39.9	126	15.8	56.2	42.3	5.55e-03	2.75e-02
PrP_0.5uM	180.2	180.1	1.9e-11	6.87	3.55	280	1.15	0.1	124	2.95e-09	1.5	3.0	3.0	8.4	3.0	71.8	49.2	159	19.5	158	40.4	9.3e-04	3.94e-03
PrP_2.4uM	180.2	180.1	1.9e-11	6.87	3.55	280	1.15	0.1	124	2.95e-09	1.5	3.0	3.0	8.4	3.0	71.8	49.2	159	19.5	158	40.4	9.3e-04	3.94e-03
PrP_24.2uM	180.2	180.1	1.9e-11	6.87	3.55	280	1.15	0.1	124	2.95e-09	1.5	3.0	3.0	8.4	3.0	71.8	49.2	159	19.5	158	40.4	9.3e-04	3.94e-03
PFOS_60nM	500.1	499.9	2 e-15	464	4.9	229	1.84	2.7	190	1.82e-11	1.3	-1.5	-3.4	4.8	6.0	51.9	51.5	272	20.4	1460	19.6	2.45e-06	4.09e-04
PFOS_0.6uM_B	500.1	499.9	2 e-15	464	4.9	229	1.84	2.7	190	1.82e-11	1.3	-1.5	-3.4	4.8	6.0	51.9	51.5	272	20.4	1460	19.6	2.45e-06	4.09e-04
PFOS_2uM	500.1	499.9	2 e-15	464	4.9	229	1.84	2.7	190	1.82e-11	1.3	-1.5	-3.4	4.8	6.0	51.9	51.5	272	20.4	1460	19.6	2.45e-06	4.09e-04
Cortisol_1uM	362.5	362.2	6.2e-11	4.37	97.7	466	1.29	0.5	307	9.12e-12	1.6	1.6	1.6	9.3	1.6	219	95.6	281	37.9	21900	55.9	3.55e-10	5.23e-04
Triptolide_1.6uM	360.4	360.2	1.1e-11	95.3	66.1	462	1.43	1.9	223	2.95e-09	1.7	3.0	2.9	10.5	1.3	193	88.3	243	35	1460	61.8	1.82e-09	1.80e-03

Continued on next page

Table B.2 – continued from previous page

Chemical Label	AM	MM	AHR	BF	BHL	BP	D	FHL	FP	HL	IOR	Log D5.5	Log D7.4	Log Koa	Log Kow	MP	MR	MV	P	Koc	ST	VP	WS
Afid_0.2nM	593.7	593.3	1.1e-10	5.01	148	602	1.39	0.3	403	8.71e-11	1.6	1.4	0.6	9.7	2.7	202	151	426	60	85100	65.9	6.44e-11	6.62e-05
Afid_1.7nM	593.7	593.3	1.1e-10	5.01	148	602	1.39	0.3	403	8.71e-11	1.6	1.4	0.6	9.7	2.7	202	151	426	60	85100	65.9	6.44e-11	6.62e-05
Broflanilide_15.1nM	663.3	662.0	1.7e-11	372	3.55	447	1.63	15.8	237	1.35e-09	1.5	6.8	6.6	10.3	5.9	143	128	408	50.6	2.4e-05	37.5	9.14e-09	3.48e-08
Broflanilide_1.5nM	663.3	662.0	1.7e-11	372	3.55	447	1.63	15.8	237	1.35e-09	1.5	6.8	6.6	10.3	5.9	143	128	408	50.6	2.4e-05	37.5	9.14e-09	3.48e-08
IFO_0.4uM	261.1	260.0	2.4e-11	2.4	4.27	336	1.32	0.1	156	1.05e-06	1.5	0.9	0.8	8.5	0.8	58.8	58.1	196	23	42.7	44.3	9.77e-05	0.0848
IFO_3.8nM	261.1	260.0	2.4e-11	2.4	4.27	336	1.32	0.1	156	1.05e-06	1.5	0.9	0.8	8.5	0.8	58.8	58.1	196	23	42.7	44.3	9.77e-05	0.0848
AVS_0.2nM	558.7	558.3	2.5e-11	7.21	85.1	694	1.27	0.2	390	1.1e-11	1.6	6.8	6.8	9.5	5.4	225	155	452	61.5	2.4e-05	46.0	9.75e-13	1.37e-06
AVS_1.8nM	558.7	558.3	2.5e-11	7.21	85.1	694	1.27	0.2	390	1.1e-11	1.6	6.8	6.8	9.5	5.4	225	155	452	61.5	2.4e-05	46.0	9.75 e-13	1.37e-06
CIT_3.1uM	324.4	324.2	1.7e-11	103	3.55	394	1.18	1.7	235	3.98e-06	1.6	0.6	2.2	10.6	3.0	122	92.1	273	36.5	2400	49.9	2.84e-08	7.73e-05
CIT_30.8nM	324.4	324.2	1.7e-11	103	3.55	394	1.18	1.7	235	3.98e-06	1.6	0.6	2.2	10.6	3.0	122	92.1	273	36.5	2400	49.9	2.84e-08	7.73e-05
EtOH_217mM	46.1	46.0	3.3e-12	1.540	4.57	70.3	0.846	0.1	9.86	5.01e-06	1.4	-0.3	-0.3	3.2	-0.2	-87.8	12.8	59.1	5.09	1.89	23.2	58.9	11.9

Continued on next page

Table B.2 – continued from previous page

Chemical Label	AM	MM	AHR	BF	BHL	BP	D	FHL	FP	HL	IOR	Log D5.5	Log D7.4	Log Koa	Log Kow	MP	MR	MV	P	Koc	ST	VP	WS
HTO_3.3uM	22.0	22.0	1.1e-11	95.3	5.75	100	0.998	1.9	190	2.95e-09	1.3	3.0	2.9	8.6	-1.4	113	3.68	18	1.46	1460	72.3	24.5	55.5

Table B.3: List of pathways and biological processes identified to be related to pancreatic development and function. The concept type refers to the database the pathway is annotated from. The degination of mapped or zebrafish is given to a pathway if the pathway has known complete zebrafish annotation (Zebrafish) or has been mapped to zebrafish genes from human orthology (Mapped).

Concept Type	**Concept ID**	**Concept Name**	**Pathway Category**	**Utilized Zebrafish or Mapped Human**
KEGG Pathway	4931	Insulin resistance	Glucoregulation	Mapped
KEGG Pathway	4911	Insulin secretion	Glucoregulation	Mapped
KEGG Pathway	4972	Pancreatic secretion	Glucoregulation	Mapped
KEGG Pathway	4930	Type II diabetes mellitus	Diabetes	Mapped
KEGG Pathway	4940	Type I diabetes mellitus	Diabetes	Mapped
KEGG Pathway	4910	Insulin signaling pathway	Diabetes	Zebrafish
KEGG Pathway	4950	Maturity onset diabetes of the young	Diabetes	Mapped
KEGG Pathway	4922	Glucagon signaling pathway	Glucoregulation	Mapped
KEGG Disease	H00409	Type 2 diabetes mellitus	Diabetes	Mapped
KEGG Disease	H00920	Exocrine pancreatic insufficiency, dyserythropoietic anemia, and calvarial hyperostosis	Digestion	Mapped
KEGG Disease	H00408	Type 1 diabetes mellitus	Diabetes	Mapped
KEGG Disease	H00932	Tropical calcific pancreatitis	Pancreatic Disease	Mapped
KEGG Disease	H00543	Renal-hepatic-pancreatic dysplasia	Development	Mapped
KEGG Disease	H00512	Permanent neonatal diabetes mellitus	Diabetes	Mapped
KEGG Disease	H00513	Transient neonatal diabetes mellitus	Diabetes	Mapped
KEGG Disease	H01224	Ketosis-prone diabetes mellitus	Diabetes	Mapped
KEGG Disease	H00942	Rabson-Mendenhall syndrome	Diabetes	Mapped
KEGG Disease	H01228	Insulin-resistant diabetes mellitus with acanthosis nigricans; Type A insulin resistance	Diabetes	Mapped
KEGG Disease	H00861	Pancreatic agenesis	Development	Mapped
KEGG Disease	H00933	Hereditary pancreatitis; Hereditary chronic pancreatitis	Pancreatic Disease	Mapped
KEGG Disease	H00410	Maturity onset diabetes of the young (MODY)	Diabetes	Mapped
KEGG Disease	H00019	Pancreatic cancer	Pancreatic Disease	Mapped
KEGG Disease	H00854	Wolfram syndrome; DIDMOAD syndrome	Diabetes	Mapped
KEGG Disease	H00045	Pancreatic neuroendocrine tumor	Pancreatic Disease	Mapped
KEGG Disease	H00766	Wolcott-Rallison syndrome	Diabetes	Mapped

Continued on next page

Table B.3 – continued from previous page

Concept Type	Concept ID	Concept Name	Pathway Category	Utilized Zebrafish or Mapped Human
KEGG Disease	H02198	Pancreatic agenesis and congenital heart disease; Pancreatic hypoplasia diabetes heart disease; Yorifuji-Okuno syndrome	Development	Mapped
GO MF	GO:0001602	pancreatic polypeptide receptor activity	Glucoregulation	Mapped
GO MF	GO:0043560	insulin receptor substrate binding	Glucoregulation	Zebrafish
GO MF	GO:0043559	insulin binding	Glucoregulation	Mapped
GO MF	GO:0005009	insulin receptor activity	Glucoregulation	Zebrafish
GO MF	GO:0005158	insulin receptor binding	Glucoregulation	Zebrafish
GO MF	GO:0004967	glucagon receptor activity	Glucoregulation	Zebrafish
GO MF	GO:0120022	glucagon family peptide binding	Glucoregulation	Mapped
GO MF	GO:0031769	glucagon receptor binding	Glucoregulation	Zebrafish
GO BP	GO:2000080	negative regulation of canonical Wnt signaling pathway involved in controlling type B pancreatic cell proliferation	Development	Mapped
GO BP	GO:0090188	negative regulation of pancreatic juice secretion	Digestion	Mapped
GO BP	GO:0031017	exocrine pancreas development	Development	Zebrafish
GO BP	GO:2000074	regulation of type B pancreatic cell development	Development	Zebrafish
GO BP	GO:0061113	pancreas morphogenesis	Development	Zebrafish
GO BP	GO:0090104	pancreatic epsilon cell differentiation	Development	Zebrafish
GO BP	GO:0003311	pancreatic D cell differentiation	Development	Zebrafish
GO BP	GO:0003326	pancreatic A cell fate commitment	Development	Mapped
GO BP	GO:0003329	pancreatic PP cell fate commitment	Development	Mapped
GO BP	GO:2000675	negative regulation of type B pancreatic cell apoptotic process	Development	Mapped
GO BP	GO:0003327	type B pancreatic cell fate commitment	Development	Mapped
GO BP	GO:0003323	type B pancreatic cell development	Development	Zebrafish
GO BP	GO:0097050	type B pancreatic cell apoptotic process	Development	Mapped
GO BP	GO:1904691	negative regulation of type B pancreatic cell proliferation	Development	Mapped
GO BP	GO:0031016	pancreas development	Development	Zebrafish
GO BP	GO:0035469	determination of pancreatic left/right asymmetry	Development	Zebrafish

Continued on next page

Table B.3 – continued from previous page

Concept Type	Concept ID	Concept Name	Pathway Category	Utilized Zebrafish or Mapped Human
GO BP	GO:2000081	positive regulation of canonical Wnt signaling pathway involved in controlling type B pancreatic cell proliferation	Development	Mapped
GO BP	GO:0003310	pancreatic A cell differentiation	Development	Zebrafish
GO BP	GO:0003309	type B pancreatic cell differentiation	Development	Zebrafish
GO BP	GO:0036321	ghrelin secretion	Digestion	Mapped
GO BP	GO:2000674	regulation of type B pancreatic cell apoptotic process	Development	Zebrafish
GO BP	GO:2000676	positive regulation of type B pancreatic cell apoptotic process	Development	Mapped
GO BP	GO:0044342	type B pancreatic cell proliferation	Development	Zebrafish
GO BP	GO:2000078	positive regulation of type B pancreatic cell development	Development	Mapped
GO BP	GO:1904692	positive regulation of type B pancreatic cell proliferation	Development	Mapped
GO BP	GO:0061469	regulation of type B pancreatic cell proliferation	Development	Zebrafish
GO BP	GO:2000227	negative regulation of pancreatic A cell differentiation	Development	Zebrafish
GO BP	GO:2000077	negative regulation of type B pancreatic cell development	Development	Zebrafish
GO BP	GO:0072560	type B pancreatic cell maturation	Development	Mapped
GO BP	GO:2000231	positive regulation of pancreatic stellate cell proliferation	Development	Mapped
GO BP	GO:0031018	endocrine pancreas development	Development	Zebrafish
GO BP	GO:0003322	pancreatic A cell development	Development	Zebrafish
GO BP	GO:1990798	pancreas regeneration	Development	Zebrafish
GO BP	GO:0030157	pancreatic juice secretion	Digestion	Zebrafish
GO BP	GO:0061114	branching involved in pancreas morphogenesis	Development	Mapped
GO BP	GO:0035773	insulin secretion involved in cellular response to glucose stimulus	Glucoregulation	Zebrafish
GO BP	GO:0032869	cellular response to insulin stimulus	Glucoregulation	Zebrafish
GO BP	GO:0032868	response to insulin	Glucoregulation	Zebrafish
GO BP	GO:0008286	insulin receptor signaling pathway	Glucoregulation	Zebrafish

Continued on next page

Table B.3 – continued from previous page

Concept Type	Concept ID	Concept Name	Pathway Category	Utilized Zebrafish or Mapped Human
GO BP	GO:0061179	negative regulation of insulin secretion involved in cellular response to glucose stimulus	Glucoregulation	Mapped
GO BP	GO:0061178	regulation of insulin secretion involved in cellular response to glucose stimulus	Glucoregulation	Zebrafish
GO BP	GO:0035774	positive regulation of insulin secretion involved in cellular response to glucose stimulus	Glucoregulation	Zebrafish
GO BP	GO:1901142	insulin metabolic process	Glucoregulation	Mapped
GO BP	GO:1901143	insulin catabolic process	Glucoregulation	Mapped
GO BP	GO:0032024	positive regulation of insulin secretion	Glucoregulation	Zebrafish
GO BP	GO:0046627	negative regulation of insulin receptor signaling pathway	Glucoregulation	Zebrafish
GO BP	GO:0046626	regulation of insulin receptor signaling pathway	Glucoregulation	Zebrafish
GO BP	GO:0046628	positive regulation of insulin receptor signaling pathway	Glucoregulation	Zebrafish
GO BP	GO:0046676	negative regulation of insulin secretion	Glucoregulation	Zebrafish
GO BP	GO:0030073	insulin secretion	Glucoregulation	Zebrafish
GO BP	GO:0030070	insulin processing	Glucoregulation	Mapped
GO BP	GO:0044381	glucose import in response to insulin stimulus	Glucoregulation	Zebrafish
GO BP	GO:1900077	negative regulation of cellular response to insulin stimulus	Glucoregulation	Mapped
GO BP	GO:1900078	positive regulation of cellular response to insulin stimulus	Glucoregulation	Mapped
GO BP	GO:1900076	regulation of cellular response to insulin stimulus	Glucoregulation	Zebrafish
GO BP	GO:0038016	insulin receptor internalization	Glucoregulation	Zebrafish
GO BP	GO:0038020	insulin receptor recycling	Glucoregulation	Mapped
GO BP	GO:0050796	regulation of insulin secretion	Glucoregulation	Zebrafish
GO BP	GO:0120116	glucagon processing	Glucoregulation	Mapped
GO BP	GO:0071377	cellular response to glucagon stimulus	Glucoregulation	Mapped
GO BP	GO:0070092	regulation of glucagon secretion	Glucoregulation	Mapped
GO BP	GO:0070093	negative regulation of glucagon secretion	Glucoregulation	Mapped
GO BP	GO:0033762	response to glucagon	Glucoregulation	Mapped
GO BP	GO:1904659	glucose transmembrane transport	Glucoregulation	Zebrafish
CO CC	GO:0032593	insulin-responsive compartment	Glucoregulation	Zebrafish
CO CC	GO:0005899	insulin receptor complex	Glucoregulation	Zebrafish

Appendix C

Supplementary Material for Chapter 4

C.1 Supplementary Tables

Table C.1: Characteristics [mean ± SD or n (%)] of subject cohorts without and with birth defects in California from 2018-2019 for each variable subset that is utilized for both multilayer network modeling and machine learning: Labor & Delivery Complications, Pregnancy Complications, Parent Giving Birth & Infant Characteristics and Physical & Social Environment. *Italicized variables* indicate categories of non-binary categorical variables.

	Without Birth Defects (N=549,675)	**With Birth Defects (N=3,013)**
Labor & Delivery Complications		
Premature Rupture of Membranes	20248 (3.68%)	121 (4.02%)
Precipitous Labor	23298 (4.24%)	63 (2.09%)
Prolonged Labor	8307 (1.51%)	28 (0.93%)
Induction of Labor	103587 (18.85%)	413 (13.71%)
Augmentation of Labor	103816 (18.89%)	291 (9.66%)
Non-vertex Presentation	17857 (3.25%)	152 (5.04%)
Steroids for Fetal Lung Maturation	13487 (2.45%)	147 (4.88%)
Antibiotics	142625 (25.95%)	531 (17.62%)
Clinical Chorioamnionitis or Maternal Temperature	12559 (2.28%)	42 (1.39%)
Meconium Staining of the Amniotic Fluid	34351 (6.25%)	143 (4.75%)
Fetal Intolerance of Labor	24591 (4.47%)	184 (6.11%)
Epidural or Spinal Anesthesia	384469 (69.94%)	2446 (81.18%)
Mother Transferred for Delivery	1109 (0.20%)	39 (1.29%)
Rupture of Membranes Prior to Onset of Labor	17315 (3.15%)	78 (2.59%)

Continued on next page

Table C.1 – continued from previous page

	Without Birth Defects (N=549,675)	With Birth Defects (N=3,013)
Abruptio Placenta	2555 (0.46%)	23 (0.76%)
Placental Insufficiency	512 (0.09%)	1 (0.03%)
Prolapsed Cord	696 (0.13%)	5 (0.17%)
Chorioamnionitis	7001 (1.27%)	25 (0.83%)
Maternal Blood Transfusion	1204 (0.22%)	10 (0.33%)
Third or Fourth Degree Perineal Laceration	3802 (0.69%)	9 (0.30%)
Ruptured Uterus	159 (0.03%)	2 (0.07%)
Unplanned Hysterectomy	248 (0.05%)	4 (0.13%)
Admission to ICU	874 (0.16%)	5 (0.17%)
Unplanned Operating Room Procedure Following Delivery	726 (0.13%)	6 (0.20%)
Other Labor/Delivery Complications	88675 (16.13%)	522 (17.32%)
Pregnancy Complications		
Prepregnancy Diabetes	4240.0 (0.77%)	41.0 (1.36%)
Gestational Diabetes	39489.0 (7.18%)	169.0 (5.61%)
Prepregnancy Hypertension	6475.0 (1.18%)	45.0 (1.49%)
Gestational Hypertension	29873.0 (5.43%)	150.0 (4.98%)
Eclampsia	460.0 (0.08%)	2.0 (0.07%)
Large fibroids	2054.0 (0.37%)	16.0 (0.53%)
Asthma	15505.0 (2.82%)	92.0 (3.05%)
Multiple pregnancy	14747.0 (2.68%)	83.0 (2.75%)
Intrauterine growth restricted birth	5948.0 (1.08%)	118.0 (3.92%)
Previous preterm birth	8184.0 (1.49%)	65.0 (2.16%)
Other previous poor pregnancy outcomes	11369.0 (2.07%)	113.0 (3.75%)
Cervical cerclage	751.0 (0.14%)	8.0 (0.27%)
Tocolysis	962.0 (0.18%)	13.0 (0.43%)
External cephalic version (Successful)	615.0 (0.11%)	8.0 (0.27%)
External cephalic version (Failed)	472.0 (0.09%)	7.0 (0.23%)
Consultation with specialist for high risk obstetric services	22066.0 (4.01%)	330.0 (10.95%)
Fertility-enhancing drugs, artificial insemination or intrauterine insemination	2323.0 (0.42%)	9.0 (0.30%)
Assisted reproductive technology	7620.0 (1.39%)	46.0 (1.53%)
Chlamydia	4378.0 (0.80%)	45.0 (1.49%)
Gonorrhea	799.0 (0.15%)	5.0 (0.17%)
Group B streptococcus	68674.0 (12.49%)	414.0 (13.74%)
Hepatitis B (acute infection or carrier)	1586.0 (0.29%)	2.0 (0.07%)
Hepatitis C	1058.0 (0.19%)	11.0 (0.37%)
Herpes simplex virus (HSV)	8825.0 (1.61%)	47.0 (1.56%)
Syphilis	918.0 (0.17%)	7.0 (0.23%)
Other Pregnancy Complications	108671.0 (19.77%)	676.0 (22.44%)
Parent Giving Birth & Infant Characteristics		
Child Sex		
Female	268311 (48.81%)	1374 (45.60%)
Male	281361 (51.19%)	1638 (54.36%)
Nonbinary	3 (0.00%)	1 (0.03%)

Continued on next page

Table C.1 – continued from previous page

	Without Birth Defects (N=549,675)	**With Birth Defects (N=3,013)**
Plurality		
Single	533188 (97.00%)	2911 (96.61%)
Twin	16119 (2.93%)	100 (3.32%)
Triplet	347 (0.06%)	2 (0.07%)
Quadruplet	21 (0.00%)	0 (0.00%)
Age	30.11 ± 5.85	29.01 ± 6.34
Race		
White	388837 (70.74%)	2648 (87.89%)
Black	32889 (5.98%)	106 (3.52%)
Other Specified	31615 (5.75%)	81 (2.69%)
Filipino	15284 (2.78%)	36 (1.19%)
Asian Japanese	2678 (0.49%)	8 (0.27%)
Asian Chinese	28249 (5.14%)	37 (1.23%)
American Indian - Native American	3225 (0.59%)	16 (0.53%)
Asian Korean	4653 (0.85%)	9 (0.30%)
Asian Specified	9528 (1.73%)	20 (0.66%)
Asian Vietnamese	8018 (1.46%)	12 (0.40%)
Asian Thai	842 (0.15%)	4 (0.13%)
Asian Laotian	726 (0.13%)	1 (0.03%)
Samoan	743 (0.14%)	2 (0.07%)
Indian	15497 (2.82%)	21 (0.70%)
Asian Hmong	2951 (0.54%)	2 (0.07%)
Pacific Islander	1447 (0.26%)	3 (0.10%)
Asian Cambodian	1879 (0.34%)	5 (0.17%)
Guamanian	229 (0.04%)	1 (0.03%)
Hawaiian	378 (0.07%)	1 (0.03%)
Eskimo	4 (0.00%)	0 (0.00%)
Aleut	3 (0.00%)	0 (0.00%)
Multirace		
Hispanic Single Race	257472 (46.84%)	2201 (73.05%)
Non-Hispanic Black	27608 (5.02%)	85 (2.82%)
Two or More - Any Hispanic Status	20577 (3.74%)	67 (2.22%)
Non-Hispanic White	152522 (27.75%)	499 (16.56%)
Non-Hispanic Asian	87348 (15.89%)	146 (4.85%)
Non-Hispanic Hawaiian - Pacific Islander	2091 (0.38%)	5 (0.17%)
Non-Hispanic American Indian - Native American	1566 (0.28%)	9 (0.30%)
Non-Hispanic Other	491 (0.09%)	1 (0.03%)
Hispanic Code		
Mexican	129214 (23.51%)	1249 (41.45%)
Not Hispanic	285335 (51.91%)	781 (25.92%)
Other Hispanic	132702 (24.14%)	976 (32.39%)
Cuban	624 (0.11%)	2 (0.07%)
Puerto Rican	1800 (0.33%)	5 (0.17%)
Education		
8th Grade or Less	19756 (3.59%)	128 (4.25%)
9th through 12th - No Diploma	46321 (8.43%)	350 (11.62%)
High School Graduate or GED	138651 (25.22%)	1139 (37.80%)
Some College Credit - No Degree	112370 (20.44%)	503 (16.69%)

Continued on next page

Table C.1 – continued from previous page

	Without Birth Defects (N=549,675)	**With Birth Defects (N=3,013)**
Associate Degree	38856 (7.07%)	306 (10.16%)
Bachelors Degree	119937 (21.82%)	392 (13.01%)
Masters Degree	55290 (10.06%)	147 (4.88%)
Doctorate or Professional Degree	18494 (3.36%)	48 (1.59%)
Women Infants & Children Program (WIC)	223120 (40.59%)	1587 (52.67%)
Number of Cigarettes Prepregnancy	0.16 ± 1.77	0.20 ± 1.83
Number of Cigarettes During Trimester 1	0.09 ± 1.16	0.13 ± 1.34
Number of Cigarettes During Trimester 2	0.06 ± 0.86	0.07 ± 0.92
Number of Cigarettes During Trimester 3	0.05 ± 0.77	0.06 ± 0.84
Weight Prepregnancy	155.08 ± 38.94	161.12 ± 38.15
Weight Delivery	183.12 ± 38.54	186.51 ± 38.94
Height	63.67 ± 2.84	63.44 ± 2.81
Prepregnancy BMI	26.88 ± 6.40	28.12 ± 6.27
APGAR Min5		
0	193 (0.04%)	7 (0.23%)
1	513 (0.09%)	30 (1.00%)
2	475 (0.09%)	19 (0.63%)
3	648 (0.12%)	19 (0.63%)
4	857 (0.16%)	20 (0.66%)
5	1700 (0.31%)	20 (0.66%)
6	3663 (0.67%)	46 (1.53%)
7	10097 (1.84%)	111 (3.68%)
8	62127 (11.30%)	364 (12.08%)
9	461892 (84.03%)	2334 (77.46%)
10	7510 (1.37%)	43 (1.43%)
APGAR Min10		
0	116 (0.02%)	4 (0.13%)
1	254 (0.05%)	26 (0.86%)
2	149 (0.03%)	14 (0.46%)
3	163 (0.03%)	8 (0.27%)
4	222 (0.04%)	7 (0.23%)
5	298 (0.05%)	13 (0.43%)
6	482 (0.09%)	10 (0.33%)
7	914 (0.17%)	17 (0.56%)
8	921 (0.17%)	9 (0.30%)
9	823 (0.15%)	7 (0.23%)
10	44 (0.01%)	0 (0.00%)
Greater Than 5	545289 (99.20%)	2898 (96.18%)
Payment Prenatal Care		
Medi-Cal-without CPSP Support Services	195229 (35.52%)	1788 (59.34%)
Private Insurance Company	283232 (51.53%)	931 (30.90%)
Other Government Programs	19194 (3.49%)	48 (1.59%)
Self-Pay	18336 (3.34%)	111 (3.68%)
No Prenatal Care	2890 (0.53%)	63 (2.09%)
Medi-Cal-with CPSP Support Services	24050 (4.38%)	66 (2.19%)
Other	6744 (1.23%)	6 (0.20%)
Birthweight (grams)	3290.89 ± 557.23	3142.44 ± 689.01
Gestation (weeks)	38.60 ± 1.91	37.96 ± 2.55

Continued on next page

Table C.1 – continued from previous page

	Without Birth Defects (N=549,675)	**With Birth Defects (N=3,013)**
Gestation (days)	269.88 ± 35.06	270.35 ± 29.62
Hearing Test		
Pending	186978 (34.02%)	649 (21.54%)
Pass (Both Ears)	248064 (45.13%)	346 (11.48%)
Test Not Available	102443 (18.64%)	1860 (61.73%)
Not Screened due to Medical Condition	5751 (1.05%)	101 (3.35%)
One Ear Fail	3851 (0.70%)	19 (0.63%)
Both Ears Fail	2048 (0.37%)	29 (0.96%)
Refused	540 (0.10%)	9 (0.30%)
Number of Children Born Alive		
1	216668 (39.42%)	1096 (36.38%)
2	176886 (32.18%)	923 (30.63%)
3	91846 (16.71%)	560 (18.59%)
4	39271 (7.14%)	258 (8.56%)
5	14883 (2.71%)	94 (3.12%)
6	5741 (1.04%)	42 (1.39%)
7	2405 (0.44%)	23 (0.76%)
8	1040 (0.19%)	6 (0.20%)
9	467 (0.08%)	7 (0.23%)
10	252 (0.05%)	3 (0.10%)
11	114 (0.02%)	0 (0.00%)
12	48 (0.01%)	0 (0.00%)
13	26 (0.00%)	0 (0.00%)
14	18 (0.00%)	1 (0.03%)
15	4 (0.00%)	0 (0.00%)
16	2 (0.00%)	0 (0.00%)
18	1 (0.00%)	0 (0.00%)
19	2 (0.00%)	0 (0.00%)
23	1 (0.00%)	0 (0.00%)
Number of Children Ever Born		
1	215434 (39.19%)	1092 (36.24%)
2	176492 (32.11%)	920 (30.53%)
3	92118 (16.76%)	562 (18.65%)
4	39791 (7.24%)	256 (8.50%)
5	15284 (2.78%)	97 (3.22%)
6	5948 (1.08%)	42 (1.39%)
7	2514 (0.46%)	25 (0.83%)
8	1077 (0.20%)	7 (0.23%)
9	514 (0.09%)	8 (0.27%)
10	269 (0.05%)	3 (0.10%)
11	122 (0.02%)	0 (0.00%)
12	50 (0.01%)	0 (0.00%)
13	27 (0.00%)	0 (0.00%)
14	21 (0.00%)	1 (0.03%)
15	6 (0.00%)	0 (0.00%)
16	4 (0.00%)	0 (0.00%)
18	1 (0.00%)	0 (0.00%)
19	2 (0.00%)	0 (0.00%)

Continued on next page

Table C.1 – continued from previous page

	Without Birth Defects (N=549,675)	**With Birth Defects (N=3,013)**
23	1 (0.00%)	0 (0.00%)
Delivery Final Route		
Vaginal - Spontaneous	352663 (64.16%)	1762 (58.48%)
Vaginal - Spontaneous - After Previous Cesarean	9127 (1.66%)	25 (0.83%)
Vaginal - Vacuum	16884 (3.07%)	109 (3.62%)
Vaginal - Vacuum - After Previous Cesarean	727 (0.13%)	3 (0.10%)
Vaginal - Forceps	1330 (0.24%)	2 (0.07%)
Vaginal - Forceps - After Previous Cesarean	138 (0.03%)	1 (0.03%)
Cesarean - Primary	64640 (11.76%)	501 (16.63%)
Cesarean - Primary - Labor Attempted	21478 (3.91%)	110 (3.65%)
Cesarean - Primary - With Vacuum	1411 (0.26%)	9 (0.30%)
Cesarean - Primary - With Vacuum - Labor Attempted	1468 (0.27%)	8 (0.27%)
Cesarean - Repeat	74293 (13.52%)	463 (15.37%)
Cesarean - Repeat - With Vacuum	2900 (0.53%)	10 (0.33%)
Cesarean - Repeat - With Vacuum - Labor Attempted	177 (0.03%)	0 (0.00%)
Cesarean - Repeat - Labor Attempted	2439 (0.44%)	10 (0.33%)
Previous Cesarean		
0	459863 (83.66%)	2501 (83.01%)
1	60623 (11.03%)	304 (10.09%)
2	21503 (3.91%)	161 (5.34%)
3	6025 (1.10%)	33 (1.10%)
4	1316 (0.24%)	13 (0.43%)
5	255 (0.05%)	1 (0.03%)
6	65 (0.01%)	0 (0.00%)
7	19 (0.00%)	0 (0.00%)
8	4 (0.00%)	0 (0.00%)
9	2 (0.00%)	0 (0.00%)
Fetal Presentation		
Cephalic	525076 (95.52%)	2816 (93.46%)
Breech	20343 (3.70%)	171 (5.68%)
Other	4256 (0.77%)	26 (0.86%)
Payment Delivery		
Medical	222643 (40.50%)	1876 (62.26%)
Private Insurance	282980 (51.48%)	919 (30.50%)
Other Government	15715 (2.86%)	51 (1.69%)
Self-pay	18274 (3.32%)	144 (4.78%)
Champus-Tricare	3059 (0.56%)	13 (0.43%)
Other	6578 (1.20%)	6 (0.20%)
Indian Health Service	87 (0.02%)	1 (0.03%)
Medically Unattended Delivery	339 (0.06%)	3 (0.10%)
Physical & Social Environment		
Ozone	0.05 ± 0.01	0.05 ± 0.01
PM2.5	10.37 ± 2.16	9.79 ± 1.89
Diesel PM	0.23 ± 0.36	0.18 ± 0.21
Drinking Water	540.47 ± 233.70	505.43 ± 176.71
Lead	50.31 ± 23.47	47.29 ± 19.60

Continued on next page

Table C.1 – continued from previous page

	Without Birth Defects (N=549,675)	**With Birth Defects (N=3,013)**
Pesticides	384.74 ± 2960.42	528.78 ± 3009.70
Toxic Release	2088.88 ± 5862.24	1062.00 ± 4774.44
Traffic	1118.74 ± 983.62	947.06 ± 826.21
Cleanup Sites	8.60 ± 16.36	9.40 ± 13.48
Groundwater Threats	17.50 ± 33.87	16.33 ± 27.87
Hazardous Waste	0.50 ± 1.71	0.37 ± 1.18
Impaired Water Bodies	3.56 ± 4.85	10.29 ± 9.32
Solid Waste	1.89 ± 3.88	3.45 ± 5.82
Asthma (prevalence rate)	55.85 ± 31.53	75.19 ± 35.59
Low Birth Weight	5.10 ± 1.44	4.80 ± 1.22
Cardiovascular Disease	14.34 ± 5.42	17.13 ± 4.76
Education (regional)	20.59 ± 15.79	26.87 ± 14.46
Linguistic Isolation	10.92 ± 9.74	17.95 ± 13.42
Poverty	35.63 ± 19.17	44.38 ± 18.95
Unemployment	7.32 ± 4.03	11.80 ± 6.27
Housing Burden	19.57 ± 8.20	19.56 ± 7.73

Table C.2: Overview of the optimized hyperparameters for each machine learning model evaluated using a grid search cross validation process. All other parameters were taken as default.

Hyperparameter	**Possible Values**	**Optimized Value**
Logistic Regression		
C	[0.1, 1, 10]	0.1
Solver	['lbfgs', 'saga']	'saga'
Penalty	['l1', 'l2', 'elasticnet]	'l1'
Max Iterations	[100, 1000]	100
Class Weight	['balanced', None]	'balanced'
Random Forest		
Number of estimators	[100, 200, 500]	500
Maximum Tree Depth	[10, 20, None]	10
Minimum Samples per Split	[2, 5, 10]	10
Minimum Samples per Leaf	[1, 3, 5]	3
Class Weight	['balanced', None]	'balanced'
XGBoost		
Maximum Depth	[6, 8]	6
Learning Rate	[0.1, 0.3]	0.1
Number of Estimators	[100, 200, 500]	100
Minimum Child Weight	[1, 5, 10]	10
Scale Position Weight	[# Negative/# Positive, 1]	1
Gamma	[0, 0.1, 0.3]	0.3

C.2 Supplementary Figures

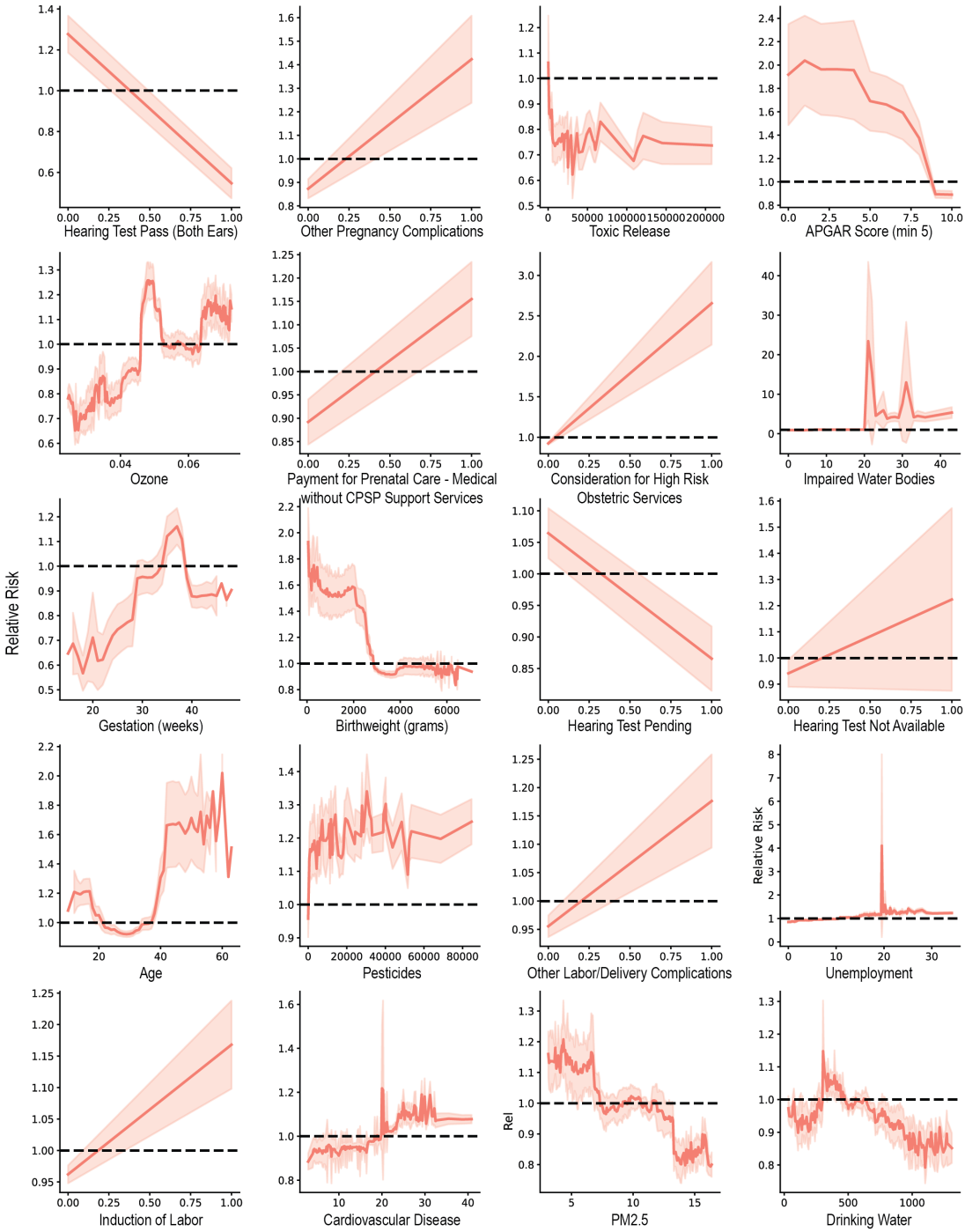

Figure C.1: Dependence plots for the top twenty features contributing to model output in the *All Variables* feature set, presented as risk ratios (RR). Shaded bands represent the standard deviation of the population per bin, and dashed lines indicate the baseline risk ratio ($RR = 1$). $RR > 1$ signifies a positive relationship with the model output, while $RR < 1$ signifies a negative relationship.

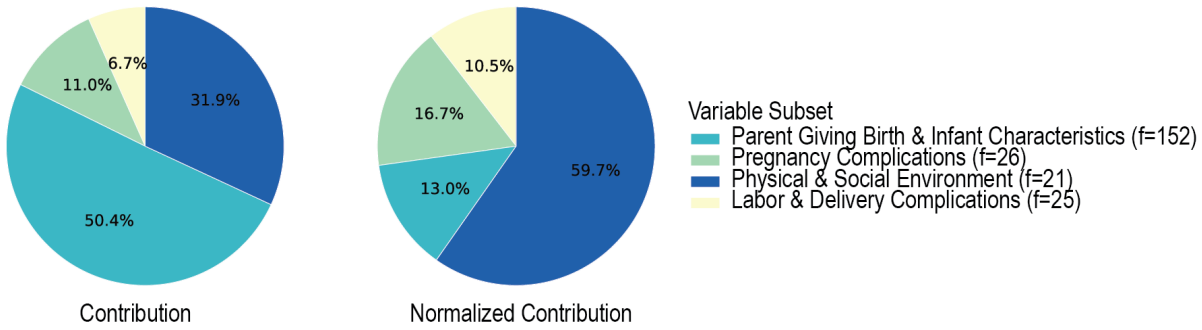

Figure C.2: Comparison of global feature importance for the *All Variables* feature set across variable subsets. (A) Original contributions based on the summed absolute SHAP values. (B) Normalized contributions, where SHAP values were summed for each variable subset and divided by the number of features (f) within the subset. The legend indicates the number of features (f) in each variable subset.

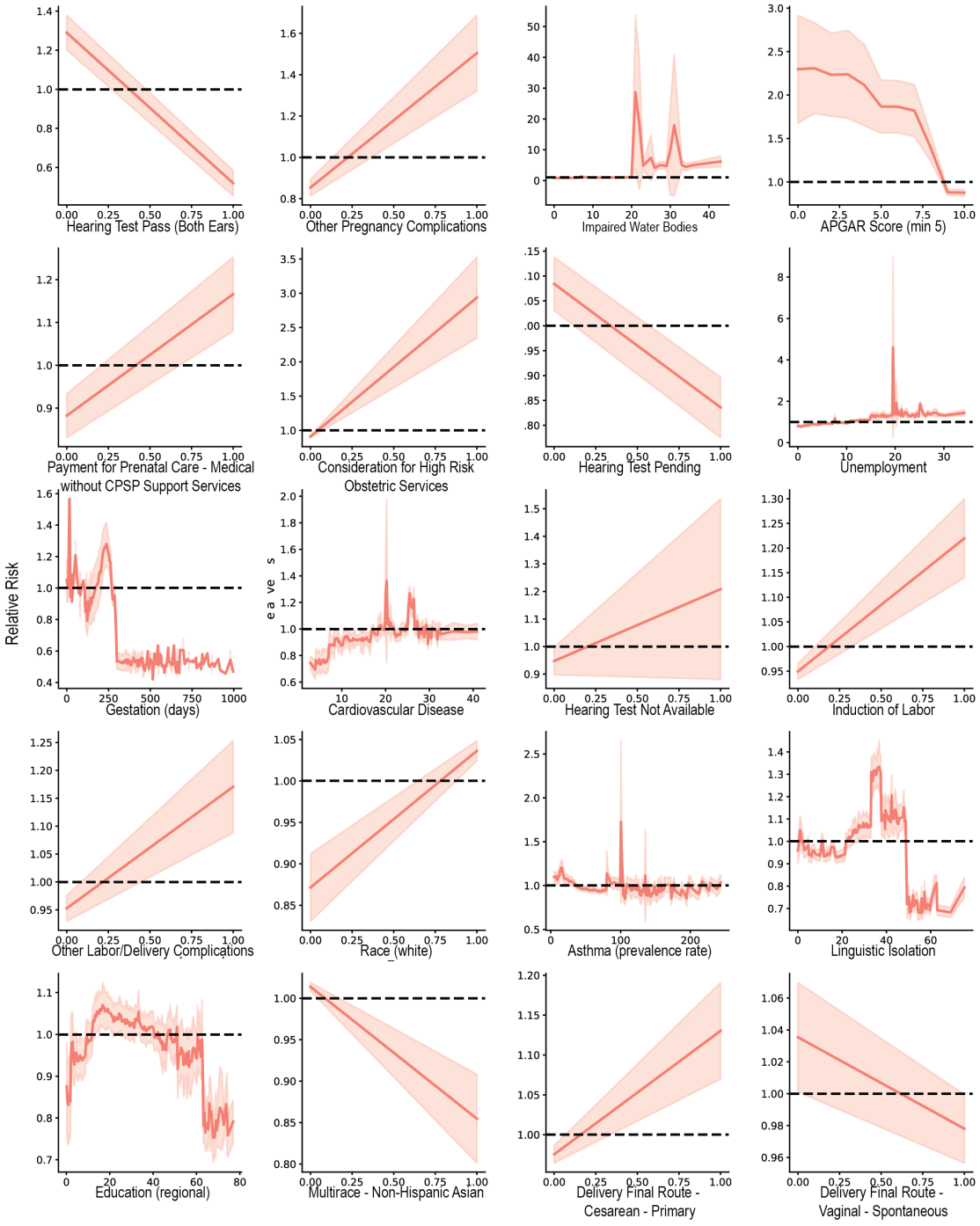

Figure C.3: Dependence plots for the top twenty features contributing to model output in the Top Variables feature set, presented as risk ratios (RR). Shaded bands represent the standard deviation of the population per bin, and dashed lines indicate the baseline risk ratio ($RR = 1$). $RR > 1$ signifies a positive relationship with the model output, while $RR < 1$ signifies a negative relationship.

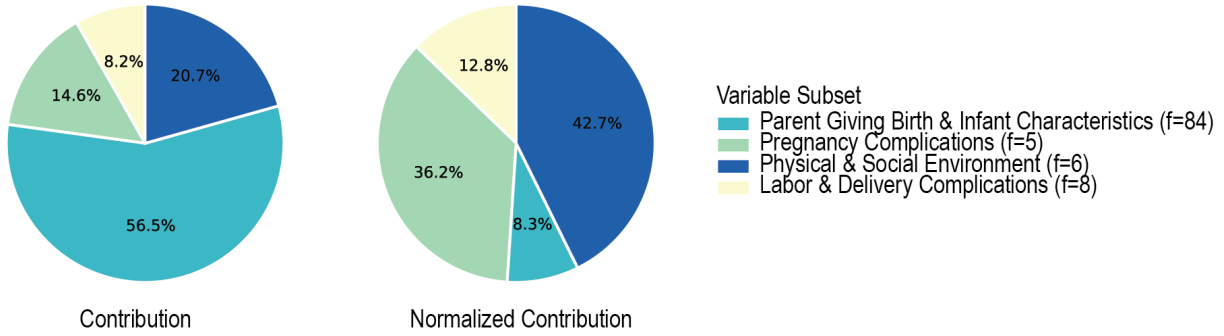

Figure C.4: Comparison of global feature importance for the Top Variables feature set across variable subsets. (A) Original contributions based on the summed absolute SHAP values. (B) Normalized contributions, where SHAP values were summed for each variable subset and divided by the number of features (f) within the subset. The legend indicates the number of features (f) in each variable subset.

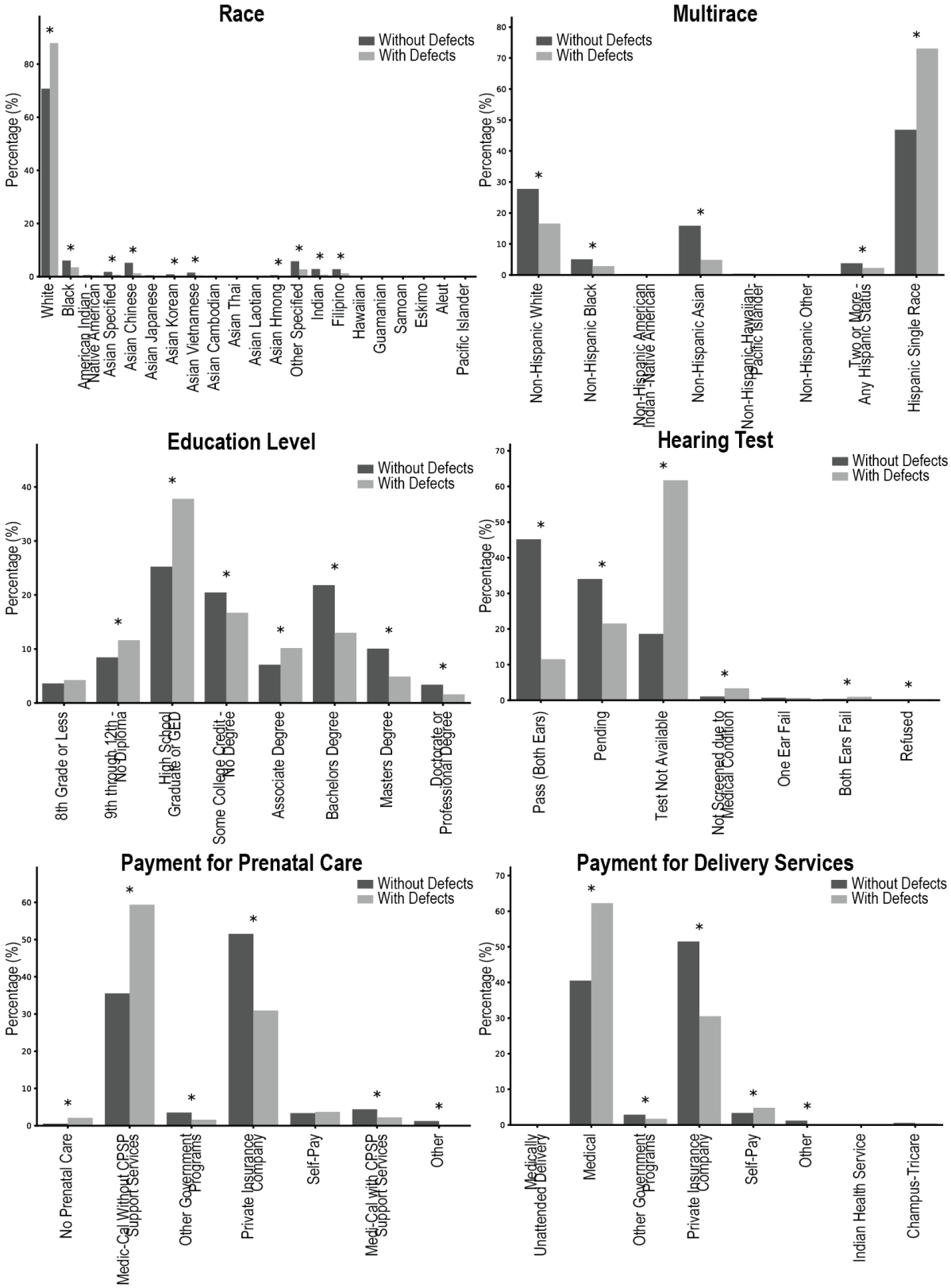

Figure C.5: Bar charts displaying the distribution of categories among top nodes with multiple categories. Statistically significant differences are marked with a black asterisk ($p < 0.05$).