

## **Distribution Agreement**

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the nonexclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

---

Zhenjiang Li

---

Date

Characterizing the Molecular Mechanisms Underlying the Health Effects of  
Fine Particulate Matters among Susceptible Populations  
via Multi-Omics Analysis

By

Zhenjiang Li  
Doctor of Philosophy  
Environmental Health Sciences

---

Donghai Liang, PhD  
Advisor

---

Anke Huels, PhD  
Co-advisor

---

Anne L. Dunlop, MD  
Committee Member

---

Alicia K. Smith, PhD  
Committee Member

---

Jeremy A. Sarnat, ScD  
Committee Member

---

Thomas S. Wingo, PhD  
Committee Member

Accepted:

---

Kimberly Jacob Arriola, Ph.D, MPH.  
Dean of the James T. Laney School of Graduate Studies

---

Date

Characterizing the Molecular Mechanisms Underlying the Health Effects of  
Fine Particulate Matters among Susceptible Populations  
via Multi-Omics Analysis

By

Zhenjiang Li  
MSPH, Emory University, 2019  
MSPH, Peking University, 2017  
MBBS, Peking University, 2015

Advisors:  
Donghai Liang, PhD  
Anke Huels, PhD

An abstract of  
A dissertation submitted to the Faculty of the  
James T. Laney School of Graduate Studies of Emory University  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy  
in Environmental Health Sciences  
2023

## Abstract

### Characterizing the Molecular Mechanisms Underlying the Health Effects of Fine Particulate Matters among Susceptible Populations via Multi-Omics Analysis

By Zhenjiang Li

In this dissertation, I explored and extended the existing advanced exposure assessment approach of fine particulate matter (PM<sub>2.5</sub>) and investigated the potential mediating role of metabolomic signatures, DNA methylation, and their interplay in the development of PM<sub>2.5</sub>-related adverse health outcomes. The dissertation applied an analytical framework that used high-dimensional mediation analysis to the analysis of metabolomic and epigenomic data (separately and simultaneously) in two cohorts, Emory Alzheimer's Disease (AD) Research Center Brain Bank and the Atlanta African American Maternal-Child Cohort, to unravel the biological mechanisms underlying the associations of PM<sub>2.5</sub> with AD-related neuropathology and preterm/early-term birth, respectively. In the brain bank cohort, we detected multiple CpG sites in prefrontal cortex tissues that mediated associations between PM<sub>2.5</sub> exposure and AD-related neuropathology markers. Some of these CpG sites are located in genes related to neuroinflammation and neuroinflammation-mediated necroptosis in brain tissues, implicating neuroinflammation a potential underlying mechanism of PM<sub>2.5</sub> neurotoxicity. In the prospective birth sample of African American pregnant people, metabolomic signatures detected in early-pregnancy serum samples were found to mediate the adverse effects of long- and short-term exposure to PM<sub>2.5</sub> on the risk of preterm birth and early-term birth (ETB). Specifically, biological pathways involved in folate metabolism and glycine and serine metabolism were found to have an important role in the biological mechanism underlying PM<sub>2.5</sub> toxicity on early birth. In addition, I employed multi-omics integration techniques and identified latent factors, derived from metabolomic and epigenomic data, that statistically mediated the association between PM<sub>2.5</sub> exposure and the risk of ETB, which provide important information about the interplay between metabolic features and DNA methylation that appears to play a critical role in PM<sub>2.5</sub> toxicity. The findings of the dissertation indicate that the use of omics techniques is valuable for capturing a holistic picture of the biological changes in human body responding to the PM<sub>2.5</sub> exposure and unravelling their role in disease development. With a focus on the vulnerable populations, the older adults and pregnant women, the current project shed light on the molecular mechanisms of the pathogenesis of ETB and AD, which is key to risk assessment and intervention development among these populations.

Characterizing the Molecular Mechanisms Underlying the Health Effects of  
Fine Particulate Matters among Susceptible Populations  
via Multi-Omics Analysis

By

Zhenjiang Li  
MSPH, Emory University, 2019  
MSPH, Peking University, 2017  
MBBS, Peking University, 2015

Advisors:  
Donghai Liang, PhD  
Anke Huels, PhD

A dissertation submitted to the Faculty of the  
James T. Laney School of Graduate Studies of Emory University  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy  
in Environmental Health Sciences  
2023

## Acknowledgement

I am deeply grateful to everyone who has supported me throughout my academic journey and helped me to complete this dissertation. It is my pleasure to acknowledge their contributions and express my heartfelt appreciation.

First and foremost, I would like to thank my advisor, Dr. Donghai Liang, for his unwavering support, guidance, and expertise. Dr. Liang's insights, mentorship, and encouragement have been instrumental in shaping my research and academic development. I am also grateful to my co-advisor, Dr. Anke Huels, for her valuable guidance and inspiration. Her expertise and insights have been invaluable in refining my research capacity.

In addition, I feel greatly indebted to be advised and supported by numerous faculty members within and outside the department at Emory: Dr. Jeremy Sarnat, Dr. Mike Caudle, Dr. Stefanie Ebel, Dr. Stefanie Eick, Dr. Anne L. Dunlop, Dr. Alicia K. Smith, Dr. Thomas S. Wingo. Their guidance and intellectual stimulation enriched my learning experience and broadened my perspectives. I acknowledge the assistance and support of the staff in the department, Angela Rozo and Ariadne Swichtenberg, whose dedication, professionalism, and kindness have made my academic journey smoother and more enjoyable.

I would like to also thank my co-workers, Ziyin Tang, Youran Tan, Qingyang Zhu, Hemali Oza, and Sarahna Moyd, for their collaboration, support, and friendship during my PhD program. Their contributions, insights, and feedback have been invaluable in shaping my research and personal growth.

My heartfelt thanks go to my wife, Xuan, for her love, patience, and support during this challenging and rewarding journey. Her unwavering support and encouragement have been a source of inspiration and motivation for me. I would like to express my sincere appreciation to my parents for their firm support throughout my academic journey. As a first-generation scholar, their love and sacrifices have been crucial in enabling me to pursue my goals and dreams.

Finally, I express my deepest gratitude to everyone who has contributed to my journey, including my family, parents, friends, colleagues, advisors, professors, collaborators, and staff. Without their support and encouragement, I would not have been able to achieve this milestone in my academic career.

## Table of Contents

Chapter 1.....	1
Introduction.....	1
Dissertation Aims .....	5
Chapter 2 <b>Application of a fusion approach to obtain spatially resolved (200m) ambient fine particulate matter in Metropolitan Atlanta</b> .....	7
Introduction.....	7
Methods .....	8
Results .....	11
Discussion.....	12
Conclusions.....	15
Tables and figures.....	16
Appendix for Chapter 2 .....	20
Chapter 3 <b>Differential DNA Methylation in the Brain as Potential Mediator of the Association between Traffic-related PM2.5 and Neuropathology Markers of Alzheimer’s Disease</b> .....	22
Background .....	22
Methods .....	24
Results .....	30
Discussion.....	32
Conclusions.....	37
Tables and figures.....	38
Appendix for Chapter 3 .....	45
Chapter 4 <b>Metabolic Signatures as Potential Mediators in the Relationship between Ambient fine particulate matter and Pre- and Early-term Births in the Atlanta African American Maternal-Child Cohort</b> .....	55
Introduction.....	55
Methods .....	56
Results .....	57

Discussion .....	65
Conclusions .....	70
Tables and figures.....	71
Appendix for Chapter 4 .....	78
<b>Chapter 5 Integration of Metabolome and Genome-wide DNA Methylation in Unravelling the Biological Mechanisms Underlying the Relationship between Ambient PM<sub>2.5</sub> and Early-term Birth</b> .....	87
Introduction.....	87
Methods .....	90
Results .....	96
Discussion .....	99
Conclusions .....	103
<b>Chapter 6 Implications and Future Direction</b> .....	112
References.....	120



# Chapter 1

## Introduction

Fine particulate matter (PM<sub>2.5</sub>) refers to the particles in the air that have an aerodynamic diameter of 2.5 micrometers or less.<sup>1</sup> PM<sub>2.5</sub> is the most studied air pollutant in previous literature and emitted from a variety of sources including mobile vehicles, industrial and agricultural activities.<sup>2</sup> In addition to the well-established health impacts of PM<sub>2.5</sub> on respiratory and cardiovascular diseases, emerging evidence in the past decade has documented that specific populations, such as pregnant people and older adults, are particularly vulnerable to the adverse effects of PM<sub>2.5</sub>, leading to a range of adverse health outcomes including adverse birth outcomes and neurodegenerative disorders.<sup>3</sup> Despite the mounting evidence underlying adverse health effects associated with exposures to PM<sub>2.5</sub>, the underlying biological mechanisms elicited from PM<sub>2.5</sub> exposures have not been fully understood. Moreover, studying the biological mechanisms of PM<sub>2.5</sub> exposure among vulnerable populations is essential, which can help in the development of targeted interventions and inform regulatory decisions related to the air quality control, such as setting specific permissible exposure limits for those populations.

It is biologically plausible that exposure to PM<sub>2.5</sub> could have adverse effects on Alzheimer's disease among the older adults as well as on birth outcomes among pregnant people. Evidence so far suggests oxidative stress and inflammation to be the major pathways underlying the toxicity of PM<sub>2.5</sub> on central neural systems.<sup>4</sup> Apart from oxidative stress and inflammation, PM<sub>2.5</sub> may also exert adverse effects on central neural system via neuronal apoptosis with synaptic damage.<sup>4</sup> PM<sub>2.5</sub> is capable of inducing the generation of free radicals, impairing mitochondrial function, and eventually disrupt the cellular balance in the amount of oxidants and antioxidants.<sup>5</sup> *In vitro*, PM<sub>2.5</sub> was shown to trigger the increase of cellular redox oxidative species (ROS) and reduction of the ratio of reduced glutathione (GSH) vs. oxidized glutathione (GSSG), and cause mitochondrial

malfunction in human neuroblastoma cell line.<sup>6</sup> The GSH/GSSG ratio has been the most widely used dynamic indicator of oxidative stress. In oligomeric beta-amyloid (A $\beta$ )-simulated microglia of mice, an elevated level of ROS was observed following PM<sub>2.5</sub> exposure, and the enhancement in oxidative stress was potentially induced by mitochondrial damage.<sup>7</sup> The presence of A $\beta$  plaques in the brain is one of the hallmarks of AD, in the deposition of which microglia plays an essential role. As for systemic inflammation, it is the major defense mechanism against external stimuli, but prolonged inflammation may be harmful. Alteration in inflammatory pathways has been proposed as an important mechanism of PM<sub>2.5</sub> toxicity.<sup>5</sup> Neuroinflammation is recognized as a possible link between PM<sub>2.5</sub> exposure and neurological disorders.<sup>4</sup> *In vivo*, Bhatt et al. found an increased level of cytokines belonging to the chemokine-chemoattractant class, A $\beta$  (an inducer of microglia activation and neuroinflammation), and amyloid-associated pathology (i.e., early Alzheimer-like changes) in the brains of mice exposed to PM<sub>2.5</sub> chronically.<sup>8</sup> Tyler et al. found that acute exposure to ultrafine particulates (UFP) induced significant changes in the expression of several cytokines, including C-C motif chemokine ligand 5, chemokine C-X-C motif ligand 1, transforming growth factor  $\beta$ , and tumor necrosis factor (TNF)  $\alpha$ , in the murine hippocampus assessed by transcription profiling.<sup>9</sup>

In addition to the health impact on Alzheimer's disease, a body of epidemiological studies has reported the imbalance between oxidants and antioxidants served as a potential promoter of several pregnancy-related disorders, including preterm birth and low birth weight<sup>10</sup>. For example, prenatal exposure to PM<sub>2.5</sub> was found to promote ROS levels and the expression of genes involved in oxidative stress response in cord blood.<sup>11</sup> An elevated expression of interleukin (IL)-4 was found in the fetal portion of the placenta in rats exposed to PM<sub>2.5</sub> before and during pregnancy;<sup>12</sup> in the mouse model, maternal exposure to PM<sub>2.5</sub> increased the levels of cytokines, including IL-1 $\beta$ , IL-6, and TNF-  $\alpha$ , in offspring.<sup>13</sup> Nachman et al. reported a monotonic positive relationship between PM<sub>2.5</sub> exposure during preconception and pregnancy and intrauterine inflammation which is a risk factor for preterm birth and low birth weight.<sup>14</sup> In sum, the mechanisms

underlying can be multifactorial, and several central biological pathways, such as oxidative stress and systemic inflammation, may drive the adverse effects of PM<sub>2.5</sub> exposure.

By uncovering the pathological processes at molecular level, we will be able to obtain a more accurate risk assessment and provide a basis for protecting the susceptible populations from the daily exposure to PM<sub>2.5</sub>. Advances in the understanding of the molecular mechanisms can lead to clinical and policy measures that can address the challenge of PM<sub>2.5</sub> for pregnant women and older adults.

Advances in omics technologies (i.e., genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics) permit the collection of high-dimensional molecular data from individuals and offers the potential to characterize the biological mechanisms comprehensively.<sup>15</sup> Environmental health researchers are increasingly leveraging omics technologies to investigate the associations of environmental stressors with disease mechanisms and phenotypes.<sup>16</sup> The use of high-throughput metabolomics has become more widespread in the field of environmental health research to detect the biological alterations that arise from chronic or acute exposure to air pollution.<sup>17</sup> The application of metabolomic and epigenetic data in the Atlanta AA Maternal-Child Cohort and ADRC brain bank, respectively, can contribute to identifying changes in metabolic or DNA methylation profiles that are associated with PTB/ETB and AD-related neuropathology markers, respectively. Moreover, the rapid generation of omics data further motivates the researchers to integrate the multi-omics data to obtain a more comprehensive view of progression of diseases resulted from the environmental contaminants.<sup>16</sup> In general, two integration strategies were employed: *posteriori* (analyzing omics data separately) and *a priori* (analyzing omics data simultaneous) integration, which are applicable to different conditions. For example, if two omics datasets are measured in the same biospecimens, both *posteriori* and *a priori* integration can be employed, which enables a higher flexibility to evaluate direct relationships between metabolites and DNA methylation and how they may relate to a phenotype. The integration of multi-omics data will contribute to evidence supporting causation by identifying the biomarkers indicative of

changes each molecular level and cross-validating the functional activity of identified biological pathways. For example, metabolomics and genome-wide DNA methylation data can provide complementary information about the underlying biological mechanisms involved in the development of diseases.<sup>18</sup> Metabolomics focuses on the study of small molecules (metabolites), while genome-wide DNA methylation examines changes in gene expression that occur without changes to the DNA sequence. DNA methylation can regulate the expression of genes involved in metabolic pathways, which, in turn, can impact the levels of metabolites produced in these pathways. Conversely, certain metabolites produced in metabolic pathways, such as those involved in one-carbon metabolism, can influence DNA methylation by providing methyl groups.<sup>19</sup> By combining these two approaches, researchers can gain a more comprehensive understanding of how genetic and environmental factors interact to affect disease development and progression.

With omics data, we will be able to examine the mediating role of metabolic perturbations and DNA methylation changes at CpG sites via high-dimensional mediation analysis. Mediation analysis can provide strong support for causal mechanisms between an exposure and outcomes.<sup>15</sup> The combination of mediation analysis and omics technologies permit the collection of high-dimensional molecular data and the identification of potential molecular intermediates lying between an exposure and an outcome, which has been readily applied in epidemiological research to increase the plausibility of causal inference.<sup>20,21</sup> Although a handful of high-dimensional mediation frameworks have been published in recent years,<sup>22-26</sup> the application in environmental health research is scarce. Recent applications of omics technologies within population-based settings present new opportunities for identifying molecular intermediates and related biological pathways linking the observed associations between environmental stressors and diseases from observational studies.<sup>16</sup> Although efforts have been made to identify the mediating role of metabolomics and epigenomics for the associations of air pollution with birth outcomes,<sup>27,28</sup> most previous studies were performed based on the meet-in-the-middle framework which was prone to biases.<sup>29</sup> Thus, the application of the high-dimensional mediation analysis is

necessary to further our knowledge on the adverse effects of traffic-related PM<sub>2.5</sub> on birth outcomes and neuropathology.

## Dissertation Aims

In this dissertation, I focused on investigating critical, functional roles of metabolome, epigenome, and their interrelationships as mediators of PM<sub>2.5</sub> toxicity on susceptible populations including pregnant women and people over 55 years old. These investigations provided novel mechanistic insights into the biological pathways by which exposure to PM<sub>2.5</sub> impacts birth outcomes and AD-related neuropathology.

**Aim 1** estimated the concentrations of ambient PM<sub>2.5</sub> from 2012 to 2019 in metropolitan Atlanta. I used a fusion approach to combine the highly-spatially-resolved land-use random forest model and the highly-temporally-resolved Community Multi-Scale Air Quality model to generate a PM<sub>2.5</sub> database with weekly averaged concentrations and a spatial resolution of 200m.

**Aim 2a** investigated how perturbations in epigenome mediated the associations between traffic-related PM<sub>2.5</sub> and AD-related neuropathology markers among an older population. I performed the Meet-in-the-Middle approach and high-dimensional mediation analyses to identify methylations at CpG sites that potentially mediated the associations.

**Aim 2b** investigated how perturbations in metabolome mediate the associations of ambient PM<sub>2.5</sub> with preterm and early-term birth in the Atlanta African American (AA) Maternal-Child cohort. I performed the Meet-in-the-Middle approach and high-dimensional mediation analyses to identify metabolic perturbations that potentially mediated the associations.

**Aim 3** optimized and applied a workflow for the integration of metabolomic and epigenomic data to identify biological mechanisms underlying the toxicity of PM<sub>2.5</sub> associated with adverse pregnancy outcomes. I optimized a workflow to incorporate existing bioinformatic tools, including *mixOmics* and *MOFA2*, for the assessment of how information from one omics level informed

another to bridge the gap from  $PM_{2.5}$  to preterm and early-term birth in the Atlanta AA Maternal-Child Cohort.

## **Chapter 2 Application of a Fusion Approach to Obtain Spatially Resolved (200m) Ambient Fine Particulate Matter in Metropolitan Atlanta**

### **Introduction**

Fine particulate matters ( $PM_{2.5}$ ), particles with aerodynamic diameters less than or equal to 2.5 micrometers, has been regulated by the National Ambient Air Quality Standards (NAAQS) as a criteria air pollutant since 1997 in the United States (U.S.), which is a significant contributor to individual exposure to air pollution.<sup>30</sup> Extensive research endeavors have been undertaken thus far, demonstrating the causal relationship of  $PM_{2.5}$  with respiratory and cardiovascular diseases, and emerging evidence in the past decade has documented the potential adverse health effects of  $PM_{2.5}$  across the life span including birth outcomes in early life and neurodegenerative disorders in late life.<sup>3</sup>

Ground monitoring, personal measurement, and air quality modeling are the major approaches for air pollution exposure assessment and have distinct advantages and disadvantages.<sup>31</sup> Ground monitoring sites, which are usually used for estimating exposures at residence, are usually spatially sparse compared the density of study population, limiting the capacity of characterizing the spatial variation in individual exposure.<sup>32</sup> Previous studies demonstrated that exposure assessment relying on ground monitoring can introduce measurement error, since the lack of a spatially resolved monitoring network can underestimate spatial variation in pollutant concentrations.<sup>33</sup> Personal measurement can obtain more accurate estimates of individual exposures. However, it usually involves the use of portable equipment and is labor intensive and costly, limiting its potential application in large-scale epidemiological studies and investigation on the long-term effects of air pollution. Air quality modeling addresses the disadvantages of ground monitoring and personal measurement. The air quality models usually generate spatially and temporally resolved concentration by developing algorithms

incorporating traffic volume, emission rate, land use, chemistry, physics, and meteorology, or aerosol optical depth. However, specific model results are often limited by either spatial resolution or an inability to capture complex chemistry and a vast array of emissions sources.

The metropolitan Atlanta area has a population of about 6.14 million people. The development of spatially and temporally resolved  $PM_{2.5}$  concentrations can help estimate the personal health burden related to  $PM_{2.5}$  exposure, facilitate the research to investigate the adverse impacts of  $PM_{2.5}$  on various health outcomes, and finally inform policy making. Local institutions in the metropolitan Atlanta area have been dedicated to generating air pollution databases containing  $PM_{2.5}$  concentrations that are both spatially and temporally resolved with the goal of satisfying the aims. The Atlanta Regional Commission (ARC) collaborating with the Georgia Environmental Protection Division developed a regionally applicable dispersion modeling methodology which allows the evaluation of traffic-related pollution at a high spatial resolution.<sup>34</sup> Dr. Russell's research group at Georgia Tech has advanced and applied air quality models over the years, such as a calibrated Research LINE source (R-LINE) dispersion model for  $PM_{2.5}$ , carbon monoxide, and nitrogen oxides,<sup>35</sup> and a fusion approach combining R-LINE model predictions and Community Multi-Scale Air Quality (CMAQ) model predictions.<sup>36</sup>

Based on the previous work, the current study aimed to develop an air quality model that can simulate weekly averaged  $PM_{2.5}$  concentrations with a spatial resolution of 200m in Metropolitan Atlanta from 2012-2019.

## Methods

This work describes the development of weekly-averaged ambient  $PM_{2.5}$  data using land-use random forest and fusion methods covering the Metropolitan Atlanta area from 2012-2019. The air quality model was done in two stages. First, we developed a land-use random forest model to estimate the annually-averaged traffic-related  $PM_{2.5}$ . We utilized high-density traffic



monitoring data, land use data, and surface PM<sub>2.5</sub> data to train a random forest model capable of accurately predicting annual traffic-related PM<sub>2.5</sub> concentrations at a spatial resolution of 200×200 meter. Then, we adopted a fusion approach that integrated the traffic-related PM<sub>2.5</sub> data and publicly available chemical transport simulations to estimate weekly-averaged PM<sub>2.5</sub> concentrations,<sup>36</sup> which incorporating comprehensive chemistry and emission sources, at a spatial resolution of 200×200 meter.

#### *Land-use random forest model*

The land-use random forest model was trained based on the PM<sub>2.5</sub> predictions of the Research LINE-source (R-LINE) dispersion model in 2015. The R-LINE predictions in Metropolitan Atlanta were requested from Atlanta Regional Commission (ARC), and the detailed process of model development can be found elsewhere.<sup>34</sup> Briefly, the R-LINE data cover a 20-county regional geography of 16,585 km<sup>2</sup> at 200 m resolution for 2015, which result in over 400,000 grid cells evenly distributed throughout the region. The model incorporates inputs, including meteorological parameters generated using AERMET<sup>37</sup> and AERMINUTE<sup>37</sup> and emission rates of PM<sub>2.5</sub> from regional roadways estimated by the ARC travel demand model<sup>38</sup> and the Motor Vehicle Emission Simulator 2014.<sup>39</sup> The annual averages of PM<sub>2.5</sub> concentrations were calculated based on the application of Stability Array method and calibrated by the regional PM<sub>2.5</sub> monitors.<sup>34</sup> Then, we specified four types of inputs for the land-use random forest model: traffic monitoring data, route inventory, land-use data, and the surface PM<sub>2.5</sub> data. Traffic monitoring data were collected by and shared through the courtesy of the Georgia Department of Transportation (GDOT).<sup>40</sup> The GDOT traffic monitoring data were based on actual measurements and contained accurate road geometry with traffic volume. We used the 2012-2019 GDOT annual average daily traffic (AADT) datasets joined to the roadway characteristics (RC) tables and AADT at each traffic monitoring site. Four input variables were estimated from the GDOT data: the shortest distance from the centroid of grid cell in R-LINE to the nearest

road, the sum of road length in each grid cell, the sum of AADT of roads in each grid cell, and the AADT measured at traffic stations within or nearest. We further classified GA roads into state highway routes, and others (i.e., public and private roads) according to the GDOT route inventory, and we estimated the aforementioned input variables by road types, respectively. The land cover data were accessed via Multi-Resolution Land Characteristics Consortium. The National Land Cover Database (NLCD) provided nationwide data on land cover at 30m resolution with 16 classes based on a modified Anderson Level II classification system.<sup>41</sup> We downloaded the NLCD data for 2011, 2013, 2016, and 2019. More details about the classification system can be found here.<sup>42</sup> We calculated the proportions of each class in every grid cell as the input variables of land use, and the land-use data were matched to temporally adjacent years. The surface PM<sub>2.5</sub> data were obtained from Atmospheric Composition Analysis Group.<sup>43</sup> The authors estimated annual ground-level PM<sub>2.5</sub> for 1998-2020 by combining Aerosol Optical Depth (AOD) with a chemical transport model, and subsequently calibrated to global ground-based observations using a Geographically Weighted Regression (GWR), as detailed in the reference.<sup>43</sup> We downloaded the annual data with a spatial resolution of 0.01×0.01 degree and join the R-LINE grid cells with the surface PM<sub>2.5</sub> data spatially. The land-use random forest model was trained based on the data in 2015 with the R package *randomForest*.<sup>44</sup> Two user-defined parameters, the number of trees and the number of variables randomly tried at each split, were determined by a balance of the running time and maximizing the out-of-bag R<sup>2</sup> value.

### *Fusion approach*

To estimate the weekly-averaged ambient PM<sub>2.5</sub> concentrations in Metropolitan Atlanta from 2012-2019, we adopted a fusion approach. The details of the approach can be found elsewhere.<sup>36</sup> Briefly, we fused PM<sub>2.5</sub> data from two models: one described above, and the Community Multi-Scale Air Quality (CMAQ) model. The CMAQ model is a chemical transport model developed by U.S. EPA,<sup>45</sup> which simulates daily air pollution concentrations, accounting

for emissions, meteorology, chemical reactions, and physical transport. The daily predictions of ambient  $PM_{2.5}$  are available at 12km resolution and calibrated by a Bayesian space-time downscaler model by regional monitoring data.<sup>46</sup> We downloaded the CMAQ data for 2012-2019 from U.S. EPA and averaged the daily concentrations to weekly estimates.<sup>47</sup> Then, we used a smooth curve fitting method to interpolate the 12km gridded outputs to 200m.<sup>48</sup> Then, we spatially matched the grid cells of the land-use random forest model with those of the CMAQ data. To simulate the fused results, we used the following equation:

$$PM_{200m} = [(CMAQ_{12km} - \overline{RLINE}_{200m})_{interpolated}] + RLINE_{200m}$$

where  $PM_{200m}$  denotes the weekly ambient  $PM_{2.5}$  concentrations for grid cells at 200m resolution;  $CMAQ_{12km}$  denotes the weekly ambient  $PM_{2.5}$  concentrations at 12km resolution derived from the CMAQ model;  $\overline{RLINE}_{200m}$  denotes the averaged traffic-related  $PM_{2.5}$  concentrations of the R-LINE grid cells matched to the corresponding CMAQ grid cells.

### *Evaluation methods*

The fusion approach predictions were compared with observations collected by ground monitoring sites of U.S. EPA Air Quality System (AQS). There were eleven monitoring sites within the 20-county area in Metropolitan Atlanta. We downloaded the daily monitoring data via the R package *RAQSAPI* and averaged the daily data to the weekly. Then, the monitoring sites were spatially matched with the grid cells of fusion models, and the Pearson correlation between  $PM_{2.5}$  predictions and measured data was calculated.

All spatial analyses were conducted by QGIS (version 3.28.2), and model development was conducted by R (version 4.0.2).

## Results

The land-use random forest model generated an annual traffic-related PM<sub>2.5</sub> dataset with the model performance quantified by out-of-bag R<sup>2</sup> (79.78%) and root-mean-square error (RMSE, 0.2373 µg/m<sup>3</sup>). Then, the fusion approach generated 416638 grids of approximate 200×200m within the 16,585-km<sup>2</sup> study area, and for each year from 2012-2019, the 53 weekly averages of ambient PM<sub>2.5</sub> concentrations were calculated for each grid (Figure 1A). According to the annual average throughout the study area (Figure S1), there was a long-term decreasing trend, but the changes over years were slight (an averaged decreasing rate of 0.43% per year). As showed by the weekly averages per year (Figure 1B), we did not observe a seasonal pattern of PM<sub>2.5</sub> concentrations. The mean of weekly averages had a range from 9.03 to 10.89 µg/m<sup>3</sup> (Figure 1A) for 2012-2019. The spatial variability was illustrated by the annual averages per grids (Figure 2). Major highways are clearly visible on spatial distribution maps of the estimates from the fusion approach, indicating the capacity of the fusion approach in capturing small-scale spatial gradients.

To evaluate the performance of the fusion approach, we collected the actual measures of ambient PM<sub>2.5</sub> concentrations from the ground monitoring sites of U.S. Environmental Protection Agency (EPA) Air Quality System (AQS). The 11 ground monitoring sites distributed unevenly throughout our study area with more in the urban areas and fewer in the suburban areas (Table 1 & Figure S2). Only four of the 11 monitoring sites had the data coverage from 2012-2019 (Table 1). We averaged the daily measures of PM<sub>2.5</sub> to obtain the weekly averaged concentrations and spatially join the monitoring sites with the closest grids of the fusion approach. At these monitoring sites, the fusion approach yielded a Pearson's correlation from 0.75 to 0.96. In other words, the estimates of the fusion approach were strongly correlated with the actual measures.

## Discussion

In the current analysis, we developed an air quality model that estimated the weekly averaged ambient  $PM_{2.5}$  concentrations with a spatial resolution of 200m in the Metropolitan Atlanta area from 2012-2019. Specifically, we employed a fusion approach to combine a spatial-resolved dispersion model (i.e., R-LINE model) and a temporal-resolved chemistry and transport model (i.e., CMAQ model), and obtained an air pollution database with a high spatial resolution and good temporal resolution. The resulting database, therefore, can capture the spatial and temporal variation of  $PM_{2.5}$  from the traffic-related and ambient sources. We also validated the model performance was validated by comparing with actual measures of ambient  $PM_{2.5}$  collected at 11 ground monitoring sites of U.S. EPA AQS in the study area, and we reached a strong correlation (overall Pearson correlation: 0.81).

Mounting epidemiological evidence has associated numerous health outcomes with  $PM_{2.5}$  exposures.<sup>3</sup> Accurate exposure assessment is indeed a crucial prerequisite for obtaining accurate epidemiological estimates.<sup>49,50</sup> The primary methods used in previous literature have been personal monitoring, ambient monitoring networks, and modelling.<sup>31</sup> While ambient monitoring networks can be too sparse that might be less capable of capture the spatial exposure variability, and personal monitoring is labor intensive and costly, the modelling approach is free of aforementioned limitations.<sup>51</sup> The fusion approach used in the current analysis provides a computationally efficient approach for estimating spatially and temporally resolved air pollutant concentrations from a myriad of emissions sources and chemical transformations. The resulting database can be applied to air pollution studies to estimate the either long-term or short-term individual exposure for participants lived in metropolitan Atlanta from 2000-2019. Any epidemiological studies conducted in metropolitan Atlanta can also leverage this  $PM_{2.5}$  database to calculate participant exposures and expand the original scope of research to evaluate the potential modifying effects of  $PM_{2.5}$  exposure.

The land-use random forest model generated the annual concentration database of traffic-related  $PM_{2.5}$ . Although it serves as an intermediate product for the ambient database, it could also be used in research investigating the long-term exposure to traffic-related  $PM_{2.5}$  in urban areas. Traffic-related  $PM_{2.5}$  is defined as emissions directly from the tailpipes of vehicles and from the wear and tear of brakes and tires.<sup>52</sup> These particles tend to contain more toxic substances, such as metals and organic carbon, than ambient  $PM_{2.5}$ .<sup>53</sup> Traffic-related  $PM_{2.5}$  concentrations are highest near busy roads and highways and can vary depending on traffic volume, speed, and composition. Ambient  $PM_{2.5}$ , on the other hand, includes particles from a variety of sources, such as power plants, industrial facilities, and wildfires, as well as particles that have been transported over long distances.<sup>54</sup> These particles tend to be smaller and have a higher concentration of sulfates, nitrates, and organic compounds than traffic-related  $PM_{2.5}$ . Ambient  $PM_{2.5}$  concentrations can vary by location and time of year and are influenced by factors such as weather conditions and regional air pollution levels.

However, the model presented in the current analysis have several limitations. First, the traffic monitoring data obtained from GDOT was collected by the traffic stations distributing throughout Georgia, and the number of these stations have been increasing over the years.<sup>40</sup> Thus, the land-use random forest model might offer more accurate predictions of traffic-related  $PM_{2.5}$  for latest years. Second, the akima interpolation can cause estimation artifacts and biases under certain circumstances. For example, akima interpolation can be sensitive to outliers in the original data, leading to estimates that are skewed away from the true values. Thirdly, the use of ambient measurement at residence as a surrogate for personal exposure inevitably leads to exposure misclassification to some extent due to failing to account for daily mobility and indoor sources of  $PM_{2.5}$ . Finally, because of data availability, the current database only covers the metropolitan Atlanta, which constrain its broader adoption. Despite these limitations, our approach has several notable strengths in addition to the spatial and temporal resolution. The land-use random forest model and fusion approach provide a fast way to obtain spatially- and

temporally-resolved  $PM_{2.5}$  database with acceptable prediction performance. The regionalized model fits the epidemiological research conducted in the local area better, and the model can lead to higher exposure variability for the study populations in the urban areas. Finally, the  $PM_{2.5}$  database has a wide temporal coverage that enables the estimation of long-term effects of  $PM_{2.5}$  on health outcomes.

## Conclusions

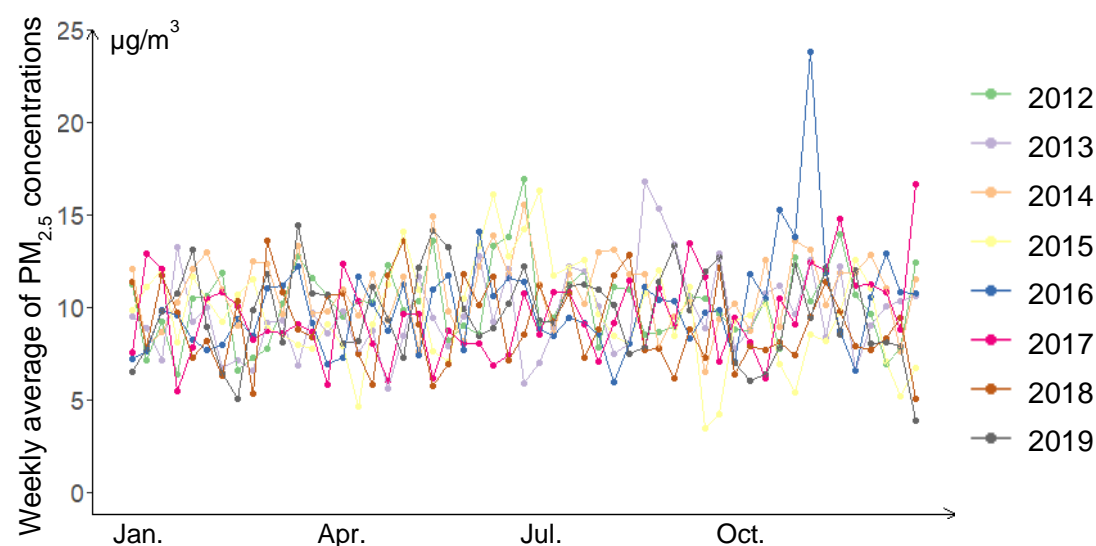
We generated a spatially- and temporally-resolved ambient  $PM_{2.5}$  database from 2012-2019 in the metropolitan Atlanta. The database can be leveraged for epidemiological studies and environmental justice studies that aim to estimate accurate individual  $PM_{2.5}$  exposures and  $PM_{2.5}$ -related health effects. The database can be applied to large-scale study populations with acceptable cost and exposure variability, which can potentially facilitate the investigation of the health effects of  $PM_{2.5}$  in diverse populations and on diverse outcomes.

## Tables and figures

### A. Statistics of weekly averaged concentrations of ambient PM<sub>2.5</sub> (µg/m<sup>3</sup>)

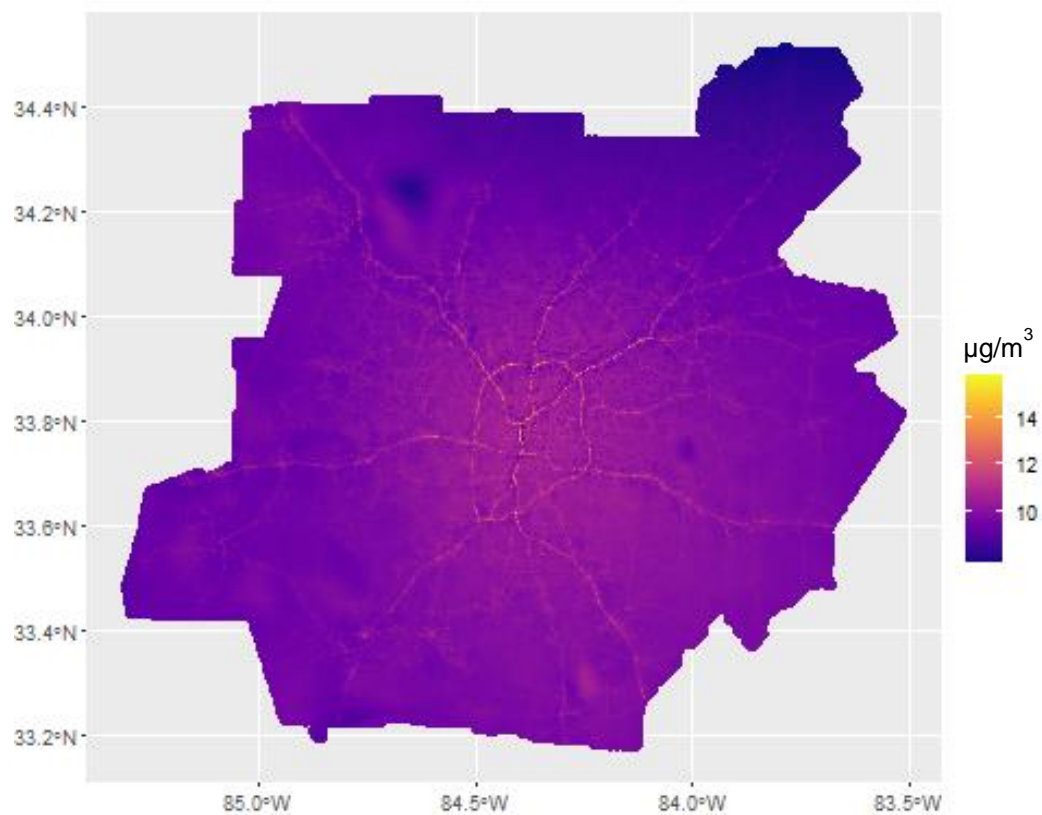
Year	N. of grids	N. of weeks	Q1	Median	Q3	IQR	Mean	SD
2012	416638	53	8.76	10.31	11.25	2.49	10.19	2.12
2013	416638	53	8.05	9.31	10.78	2.73	9.66	2.34
2014	416638	53	9.48	11.01	12.19	2.71	10.89	1.99
2015	416638	53	8.01	9.66	11.49	3.49	9.73	2.76
2016	416638	53	8.32	9.81	11.15	2.82	10.06	2.78
2017	416638	53	8.04	9.09	10.85	2.81	9.55	2.30
2018	416638	53	7.48	8.78	10.76	3.28	9.03	2.17
2019	416638	53	8.01	9.79	11.36	3.36	9.68	2.37

### B. Weekly averaged concentrations of ambient PM<sub>2.5</sub> across Metropolitan Atlanta by years



**Figure 1.** Overview of the weekly average of ambient PM<sub>2.5</sub> concentrations per year estimated by the Fusion approach from 2012 to 2019. A. Weekly averaged concentrations of ambient PM<sub>2.5</sub> across Metropolitan Atlanta by years. B. Statistics of weekly averaged concentrations of ambient PM<sub>2.5</sub>. Abbreviations: Q1, the first quartile; Q3, the third quartile; IQR, interquartile range; SD, standard deviation.

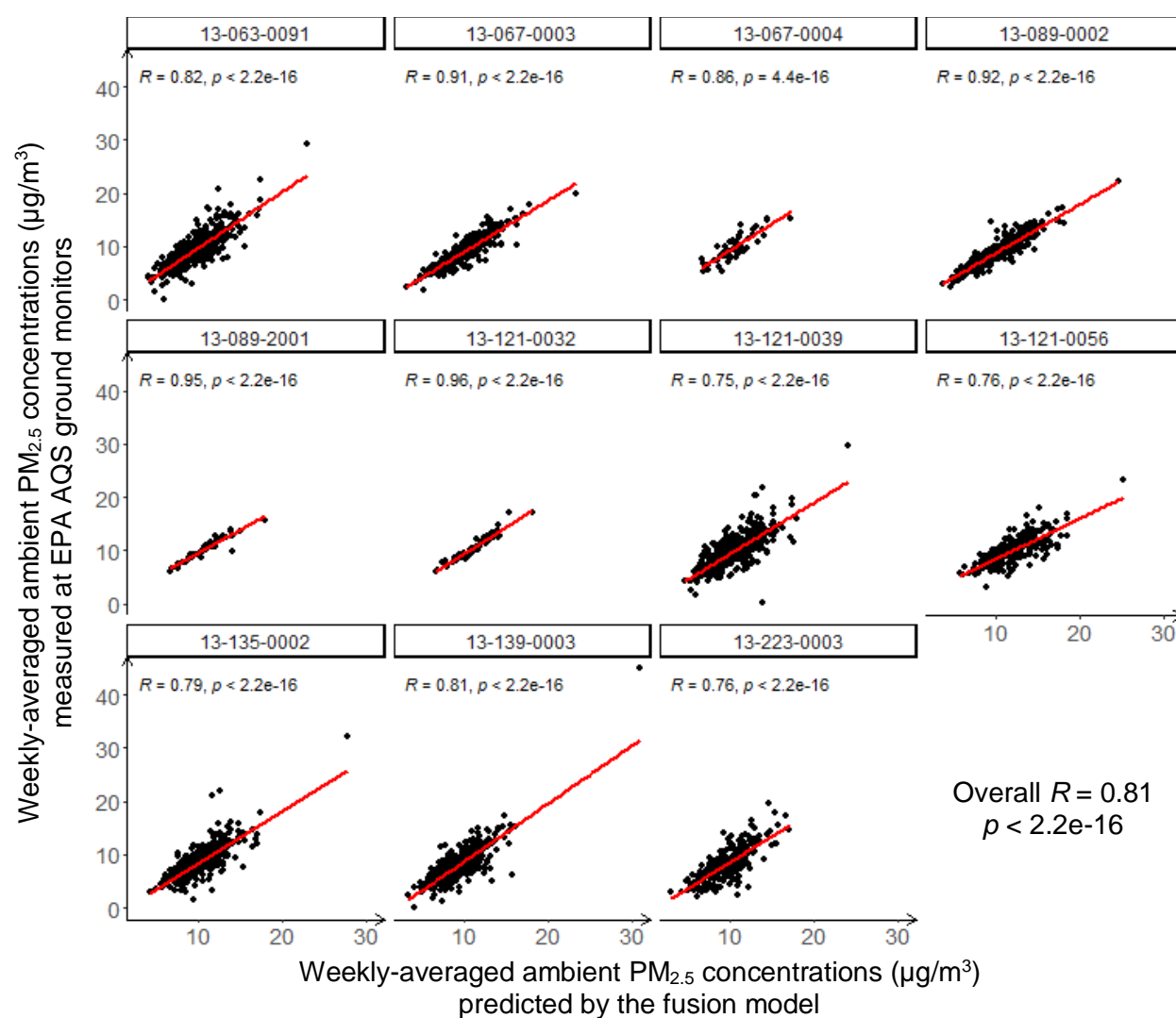




**Figure 2.** Annual averages of ambient PM<sub>2.5</sub> concentrations of 2015 estimated by the fusion approach in the Metropolitan Atlanta.

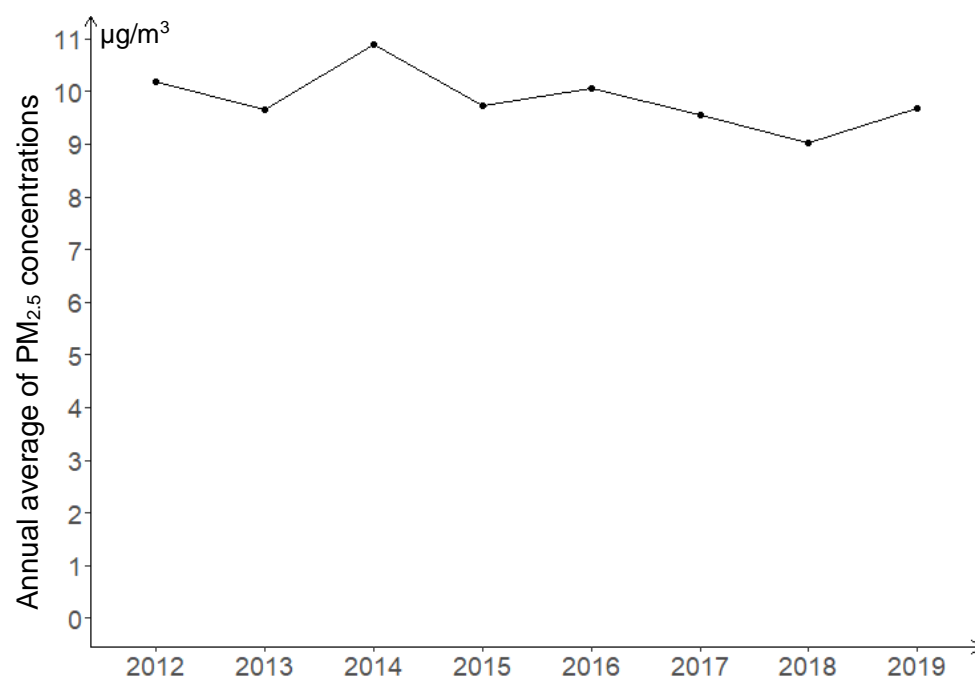
**Table 1.** Ground monitoring sites of U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) within the study area.

Site ID	County	City	Local site name	Data coverage
13-063-0091	Clayton	Forest Park	Forest Park	2012-2019
13-067-0003	Cobb	Kennesaw	Kennesaw	2012-2019
13-067-0004	Cobb	Powder Springs	Macland Aquatic	2012
13-089-0002	DeKalb	Not in a city	South DeKalb	2012-2019
13-089-2001	DeKalb	Doraville	Doraville	2012
13-121-0032	Fulton	Atlanta	E. Rivers School	2012
13-121-0039	Fulton	Atlanta	Fire Station #8	2012-2019
13-121-0056	Fulton	Atlanta	NR-GA Tech	2015-2019
13-135-0002	Gwinnett	Not in a city	Gwinnett	2012-2018
13-139-0003	Hall	Gainesville	Gainesville	2012-2018
13-223-0003	Paulding	Not in a city	Yorkville, King	2012-2017

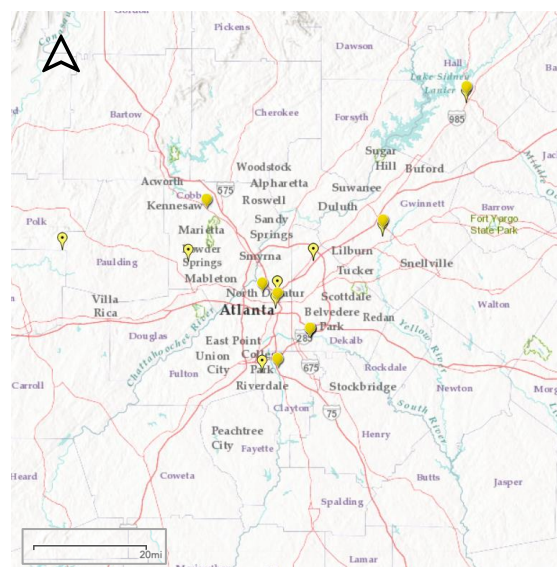


**Figure 3.** The correlation of weekly-averaged ambient  $PM_{2.5}$  concentrations between measures at ground monitors of U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) and predictions estimated by the fusion model over the study period (2012-2019).  $R$  denotes the Pearson's correlation coefficient. Each scatter plot represents the results for a ground monitoring site within the study area, which is labelled with the primary identifier for the site (site ID).

## Appendix for Chapter 2



**Figure S1.** Annual average of PM<sub>2.5</sub> concentrations calculated from the PM<sub>2.5</sub> predictions of the fusion approach in the study area.



**Figure S2.** The distribution of Ground monitoring sites of U.S. Environmental Protection Agency (EPA) Air Quality System (AQS). The yellow drop pins denote the locations of monitoring sites.

## Chapter 3 Differential DNA Methylation in the Brain as Potential Mediator of the Association between Traffic-related PM<sub>2.5</sub> and Neuropathology Markers of Alzheimer's Disease

### Background

Exposure to traffic-related air pollution (TRAP) is a significant contributor to public health burden with various well-characterized and emerging detrimental health effects.<sup>3</sup> Fine particulate matter (PM<sub>2.5</sub>), which has been regulated by the National Ambient Air Quality Standards (NAAQS) as a criteria air pollutant since 1997 in the United States (U.S.),<sup>30</sup> is an important component of TRAP mainly resulting from tailpipe exhaust, brake wear, tire wear, and resuspended dust.<sup>55</sup> A previous study has demonstrated PM<sub>2.5</sub> from traffic emissions has higher toxicity compared to other natural sources in terms of oxidative potential, cell viability, genotoxicity, oxidative stress, and inflammatory response.<sup>56</sup> The literature to date demonstrates that exposure to PM<sub>2.5</sub> is associated with a series of neurological disorders, including dementia and Alzheimer's disease (AD).<sup>57,58</sup>

AD is the most common cause of dementia and its hallmark pathologies include accumulation of beta-amyloid (A $\beta$  plaques) outside neurons and aggregation of hyperphosphorylated tau protein (neurofibrillary tangle, NFT) inside neurons in the brain.<sup>59</sup> In the U.S., 9.30 and 75.68 million people are estimated to develop clinical AD or preclinical AD by 2060,<sup>60</sup> and the total direct medical costs of AD at the national level is estimated to reach \$259 billion by 2040.<sup>61</sup> Due to the growing public concern with these substantial increases in the prevalence of AD, investigations on interventions to prevent progression and onset of AD have targeted the potentially modifiable risk factors of AD, including air pollution.<sup>62</sup>

Different biological pathways have been discussed underlying the association between air pollution and AD development. PM<sub>2.5</sub> exposure might directly infiltrate the brain<sup>63</sup> and accelerate AD pathogenesis and development via neuroinflammation, oxidative stress, and A $\beta$

accumulation.<sup>64</sup> Increasing evidence from human and animal studies proposes that perturbations in DNA methylation (DNAm), which regulate the expression of genes, are associated with indicators of AD as well as PM<sub>2.5</sub> exposure. However, the tissue specificity of DNAm has limited the ability of previous studies to formally investigate mediation.

While there is no conclusive evidence of an association between AD and DNAm in blood,<sup>65</sup> a growing body of evidence suggests robust association in brain tissues.<sup>65</sup> DNAm alterations in a number of genes were observed to be associated with AD pathology and neuroinflammation in brain tissues, such as amyloid precursor protein (*APP*),<sup>66</sup> microtubule-associated protein tau (*MAPT*),<sup>66</sup> apolipoprotein (*APOE*) promoter region,<sup>67</sup> homeobox A3 (*HOXA3*),<sup>68</sup> interleukin-1 beta (*IL-1β*),<sup>69</sup> interleukin-6 (*IL-6*),<sup>69</sup> and claudin-5 (*CLDN5*) genes.<sup>70</sup>

The association of PM<sub>2.5</sub> with DNAm in blood has been extensively studied<sup>71</sup>, and one study found that DNAm in interleukin-10 (*IL-10*), *IL-6*, tumor necrosis factor (*TNF*), toll like receptor 2 (*TLR2*) genes, which play key roles in neuroinflammation,<sup>72</sup> was significantly altered in response to short-term exposure to PM<sub>2.5</sub> and its species.<sup>73</sup> However, to the best of our knowledge, no human studies have been published on the association between PM<sub>2.5</sub> exposure and DNAm in the brain, which is the most relevant tissue when studying AD. The only evidence to date comes from *in-vivo* and *in-vitro* studies. Tachibana et al. demonstrated with a mouse model that prenatal exposure to diesel exhaust altered DNAm in brain tissues collected from 1- and 21-day-old offspring, and the differentially methylated CpG sites were enriched in the gene ontology (GO) terms related to neuronal development.<sup>74</sup> Wei et al. exposed human neuroblastoma cells to PM<sub>2.5</sub> collected at a near-road site and found that DNAm was hypermethylated in the promoter regions of neurexin 1 (*NRXN1*) and neuroligin 3 (*NLGN3*) genes encoding synaptic neuronal adhesion molecules that mediate essential signaling at synapse.<sup>75</sup>

Given the limited evidence of an association between PM<sub>2.5</sub> exposure and DNAm in the brain, the mediating role of DNAm for the association between PM<sub>2.5</sub> and AD pathology has not

been well studied. Only one study investigated DNAm in mouse brain; these investigators failed to find evidence for DNAm as a potential mediator of the association between particulate matter exposure and increased cytokines and A $\beta$  levels associated with early AD-like pathology.<sup>8</sup>

The current study investigated the relationship among PM<sub>2.5</sub>, DNAm and AD neuropathology in the post-mortem human brain among the brain donors of the Emory Goizueta AD Research Center (ADRC) brain bank. We recently showed a significant association between traffic-related PM<sub>2.5</sub> exposure and increased AD neuropathology in this dataset (Christensen et al. 2023). To elucidate the biological mechanisms for this association, we here investigated whether differential DNAm in the prefrontal cortex tissues mediates the association between long-term exposure to traffic-related PM<sub>2.5</sub> and the levels of AD-related neuropathological markers. This hypothesis was tested using a combination of the Meet-in-the-Middle (MITM) approach and high-dimensional mediation analysis.

## Methods

### *Study design*

The current cross-sectional analysis included study participants recruited by the Emory Goizueta ADRC. The ADRC was founded in 2005 and has maintained a brain bank to facilitate AD research. The study participants were research participants evaluated annually, and others were patients treated by Emory Department of Neurology physicians and diagnosed clinically with AD (biomarker defined) or probable AD. The prefrontal cortex tissues were obtained from the participants who had consented to donate biospecimens to the ADRC brain bank. There were 1011 donors enrolled by the third quarter of 2020. After applying the following inclusion criteria, 187 donors remained eligible for the current study: 1) the availability of residential addresses within Georgia (GA) state; 2) age at death equal to or over 55 years; 3) deceased after 2007; 4) no missing values in neuropathology outcomes and key covariates including age



at death, race, sex, educational attainment, and APOE genotype. Among these donors, genome-wide DNAm was measured in 161 available samples, and after quality control, 159 were included in the current analysis. Written informed consent was provided for all donors, and samples were obtained following research protocols approved by the Emory University Institutional Review Board.

### *Neuropathology assessment*

The ADRC performed thorough neuropathologic evaluations on the brains of all donors using established comprehensive research evaluations and diagnostic criteria.<sup>76</sup> These neuropathological assessments include a variety of stains and immunohistochemical preparations, as well as semi-quantitative scoring of multiple neuropathologic changes by experienced neuropathologists using published criteria.<sup>77</sup> In this project, AD-related neuropathological changes were evaluated using Braak stage, Consortium to Establish a Registry for AD (CERAD) score, and a combination of Amyloid, Braak stage, and CERAD (ABC) score which were developed based on the A $\beta$  plaques and NFTs.<sup>78</sup> Braak stage is a staging scheme describing NFTs with six stages (Stage I-VI) with a higher stage indicating a wider distribution of NFTs in brain. CERAD score describes the prevalence of A $\beta$  plaques with four levels from no neuritic plaques to frequent. ABC score combines the former two (along with the Thal score for A $\beta$  plaque distribution across various brain regions)<sup>79</sup> and is transformed into one of four levels: not, low, intermediate, or high level of AD neuropathologic changes.

### *Air pollution assessment*

Annual concentrations of traffic-related PM<sub>2.5</sub> were estimated for the 20-county area of Metropolitan Atlanta, GA for 2002-2019. The spatial resolution of the PM<sub>2.5</sub> data were 250×250m (for 2002-2011) and 200×200m (for 2012-2019). The grid cells of the corresponding side length were evenly distributed throughout the study area. The process for estimating 2002-

2011 PM<sub>2.5</sub> concentrations was previously published.<sup>1,35</sup> Briefly, a calibrated Research LINE-source dispersion (R-LINE) model for near surface releases was applied for calculating annual averages of traffic-related PM<sub>2.5</sub>. The model yielded a normalized root mean square error of 24% and a normalized mean bias of 0.3% by comparing with the estimates of the receptor-based source apportionment Chemical Mass Balance Method with Gas Constraints.<sup>35</sup> For estimating 2012 to 2019 PM<sub>2.5</sub> concentrations, we trained a land-use random forest model based on the 2015 annual concentrations of traffic-related PM<sub>2.5</sub> obtained from Atlanta Regional Commission,<sup>34</sup> road inventory and traffic monitoring data shared by the Georgia Department of Transportation,<sup>80</sup> land cover data accessed via the National Land Cover Database,<sup>81</sup> and ambient PM<sub>2.5</sub> data obtained from Atmospheric Composition Analysis.<sup>82</sup> The random forest model was trained with the R package *randomForest*<sup>83</sup>, and two user-defined parameters (i.e., the number of trees and the number of variables randomly tried at each split) were determined by a balance of the efficiency and the out-of-bag R<sup>2</sup> value. The final model reached an out-of-bag R<sup>2</sup> of 0.8 and a root-mean-square deviation of 0.2 µg/m<sup>3</sup>. This model was used to predict annual traffic-related PM<sub>2.5</sub> for 2012-2019 with a spatial resolution of 200m. Finally, we spatially matched geocoded residential addresses to the centroid of closest grids and calculated the individual long-term exposures as the average of specific exposure windows (1 year, 3 years, and 5 years prior to death).

### *Genome-wide DNA methylation*

DNA was isolated from fresh frozen prefrontal cortex in 161 samples using the QIAGEN GenePure kit. DNAm was assessed with the Illumina Infinium MethylationEPIC BeadChips in batches of 167 prefrontal cortex samples including 6 replicates. The raw intensity files were transformed into a dataset that included beta values for each the CpG sites, and these beta values were computed as the ratio of the methylated signal to the sum of the methylated and unmethylated signals, which ranged from 0 to 1 on a continuous scale. Pre-processing and

statistics were done using R (v4.2.0). We followed a validated quality control and normalization pipeline as previously published.<sup>84</sup> The detailed data processing and sample quality control can be found in the Supplementary Methods. One hundred and nine samples passed the quality check, and after excluding SNP probes, XY probes and other low-quality probes, 789,286 CpG sites remained for analysis. The final DNAm beta values were further normalized to reduce the probe type differences and corrected by ComBat to remove the batch effect before the downstream analysis.<sup>85</sup> We estimated the cell-type proportions (neuronal vs. non-neuronal cells) for each sample using the most recent prefrontal cortex database and the R package *minfi*.<sup>86,87</sup>

### *Covariate assessment*

The confounding structure was determined according to literature review and our previous studies, which was illustrated by directed acyclic graphs (DAGs) in the Supplement (Figure S1). Individual-level demographic characteristics [sex, race (Black vs. White), educational attainment (high school or less, college degree, and graduate degree), age at death, APOE  $\epsilon$ 4 genotype] were obtained from the medical records. APOE  $\epsilon$ 4 genotype was continuous with a 3-point scale (0 = no  $\epsilon$  allele, 1 = one  $\epsilon$ 4 allele, and 2 = two  $\epsilon$ 4 alleles). Area Deprivation Index (ADI) for each donor was estimated at the residential address as a proxy for neighborhood socioeconomic status, based on a publicly available database at the level of the Census Block Group for 2015.<sup>88</sup> Post-mortem interval (hours) of sample collection was provided by our lab collaborators.

### *Statistical analysis*

Previously, we found higher residential PM<sub>2.5</sub> exposure was associated with increased AD neuropathology in the Emory Goizueta ADRC brain bank (Christensen et al. 2023). To identify DNAm patterns in brain tissue that potentially mediate the association between PM<sub>2.5</sub> exposure and increased neuropathology markers, we 1) conducted an epigenome-wide association study (EWAS) for the long-term PM<sub>2.5</sub> exposures 1 year, 3 years, and 5 years prior to death and then

investigated whether any differentially methylated CpG sites that were significantly associated with PM<sub>2.5</sub> exposure in the EWAS were also associated with increased neuropathology markers; and 2) conducted a combination of Meet-in-the-Middle (MITM) approach and high-dimensional mediation analysis (HDMA) to identify any mediating CpGs that did not reach genome-wide significance in the EWAS of PM<sub>2.5</sub>. The MITM approach and HDMA work complementarily to maximize the detection of potential mediators.

Firstly, we conducted an EWAS to assess associations of long-term PM<sub>2.5</sub> exposures 1 year, 3 years, and 5 years prior to death and methylation levels of CpG sites. Specifically, we used robust multiple linear regression models as implemented in the R package *MASS* to identify differentially CpG sites associated with PM<sub>2.5</sub> exposures.<sup>89</sup> To account for measured confounding factors, we included sex, race, educational attainment, age at death, PMI, ADI, and proportion of neuronal cells in the model. Potential batch effect and other unwanted variation were further corrected using the R packages *sva*<sup>90</sup> (estimating surrogate variables included in the EWAS model as covariates) and *Bacon*.<sup>91</sup> To account for multiple testing, the *Bonferroni* threshold was used for statistical significance ( $0.05 / 789,286 = 6.33 \times 10^{-8}$ ).<sup>92</sup>

Any CpG sites that were significantly associated with PM<sub>2.5</sub> exposure were then investigated for their associations with neuropathology markers. These associations were extracted from an EWAS of each neuropathology marker (CERAD, Braak stage, ABC score) with methylation levels of all CpG sites, using robust multiple linear regression models with the neuropathology markers converted to continuous outcomes and DNAm beta values of CpG sites as exposures, adjusting for sex, race, educational attainment, age at death, PMI, APOE genotype, and proportion of neuronal cells. We used *Bacon*<sup>91</sup> to control for unmeasured confounding and bias due to the minor inflation/deflation indicated by raw *p*-values.

For the MITM, we compared the 1,000 most significant CpGs from the two sets of EWAS on all CpG sites for PM<sub>2.5</sub> exposures and neuropathology markers to identify the differentially

methyated CpG sites that were associated with both exposures and outcomes. The MITM approach is widely used in high-dimensional setting to identify intermediate biomarkers.<sup>93</sup>

Then, we conducted a HDMA using the R packages *HIMA* and *DACT* to identify any potential mediating CpG sites between PM<sub>2.5</sub> exposure and neuropathology. *HIMA* is an R package for estimating and testing high-dimensional mediation effects for omics data, which adopts the multiple mediator model's framework with reducing the dimensionality of omics data via sure independence screening and minimax concave penalty.<sup>94</sup> The divide-aggregate composite null test (*DACT*) is a more recent method for HDMA, which utilizes the Efron empirical null framework to calculate a weighted sum of *p*-values obtained from exposure-mediator (EWAS of PM<sub>2.5</sub> exposure as described above) and mediator-outcome (EWAS of neuropathology markers as described above) models for testing the significance of all mediators<sup>95</sup>. After running *HIMA* and *DACT* for all CpG sites, we further corrected for multiple testing using the *Bonferroni* method. Lastly, we combined the mediating CpG sites identified by either *HIMA* or *DACT* and used the R package *mediation* to conduct a causal mediation analysis for them to obtain the indirect effects.<sup>96-98</sup> The average causal mediation effect (i.e., indirect effect) and total effect estimated by *mediation* were summarized for the CpG sites with positive indirect effects that were in line with the hypothesized adverse effect of traffic-related PM<sub>2.5</sub> on neuropathology markers. In contrast to the MITM approach described earlier, HDMA examine multiple mediators together in a framework of mediation analysis, which allowed us to ascertain the extent to which the particular indirect effects were associated with the mediators.

To aid the interpretation of model results, we conducted a gene ontology analysis using the R package *missMethyl* based on the top 1000 CpG sites with lowest raw *p*-values<sup>99</sup>. The gene ontology analysis was conducted for the EWAS results of PM<sub>2.5</sub> exposure as well as for the EWAS results of the three neuropathology markers. All CpG sites were annotated using an online annotation data for the 'IlluminaHumanMethylationEPIC'.<sup>100</sup> Additional functional insight

on single CpG sites was obtained by searching the corresponding CpG site in publicly available databases, including EWAS catalog<sup>101</sup> and GoDMC.<sup>102</sup>

All analyses were completed in R (v4.2.0).

## Results

### *Study population characteristics*

A total of 159 donors were included in the current analysis, and their demographic characteristics and neuropathologic markers are described in Table 1. The average age of death was 76.6 years (SD=9.98) and 56% of the study population were male. The study population was predominantly white (89.3%) and well-educated with 123 (78.7%) completing college or more and living in less deprived neighborhoods (ADI: mean = 36.3, SD = 24.2). The majority of study sample (95.6%) were diagnosed with AD or other forms of dementia, and the prevalence of the APOE ε4 allele (56% with at least one APOE ε4 allele) in this population was much higher than that in the general population in the U.S.<sup>103</sup>

As illustrated by the 1-year traffic-related PM<sub>2.5</sub> exposure (Figure 1A), donors living in urban areas had a higher level of PM<sub>2.5</sub> exposure compared to those living in suburban areas. The median of 1-year exposure was 1.21 µg/m<sup>3</sup> [interquartile range (IQR)=0.78]. As PM<sub>2.5</sub> concentrations have decreased over the last decades, 3-year and 5-year exposures were slightly higher (3-year exposure: median=1.32 µg/m<sup>3</sup> [IQR=0.74], 5-year exposure: median=1.39 µg/m<sup>3</sup> [IQR: 0.81]) (Figure 1B).

### *Association between PM<sub>2.5</sub> exposure and DNAm in the brain*

After correcting for multiple tests and adjusting for bias and measured and unmeasured confounding, two CpG sites (cg25433380 and cg10495669) were consistently associated with PM<sub>2.5</sub> across different exposure windows (Figure 2, Table 2; summary statistics for all 789,286

CpG sites are provided as Table S4-6 in spreadsheets). For example, a 1  $\mu\text{g}/\text{m}^3$  increase in 1-year  $\text{PM}_{2.5}$  exposure was associated with 0.0065 increase in the DNAm beta value of cg25433380 ( $p = 1.58 \times 10^{-8}$ ). cg25433380 and cg10495669 are on chromosome 9 and 20, respectively, and cg10495669 is assigned to the gene encoding RanBP-type and C3HC4-type zinc finger-containing protein 1 (*RBCK1*). The two CpG sites were not significantly associated with any neuropathology markers (Table 2).

#### *Meet-in-the-Middle approach and high-dimensional mediation analysis*

To identify CpG sites that mediate the association between  $\text{PM}_{2.5}$  exposure and neuropathology makers but did not reach genome-wide significance in the EWAS of  $\text{PM}_{2.5}$  exposure, we conducted the MITM approach and HDMA (*DACT* and *HIMA*).

For the MITM approach, we explored the overlapping CpG sites among the top 1000 CpG sites for the EWAS of  $\text{PM}_{2.5}$  and the EWAS of neuropathology markers and identified four overlapping CpG sites (Table S1). Specifically, DNAm in cg01835635 (*APOA4* gene) was associated with CERAD score as well as  $\text{PM}_{2.5}$  exposure for the 1-year and 3-year exposure windows. DNAm in cg09830308 (*MLKL* gene) was associated with Braak stage as well as  $\text{PM}_{2.5}$  exposures for the 1-year, 3-year, and 5-year windows; cg16342341 (*SORBS2* gene) was associated with CERAD score as well as 1-year  $\text{PM}_{2.5}$  exposure; and cg27459981 (*MLKL* gene) was associated with Braak stage and ABC score as well as  $\text{PM}_{2.5}$  exposures for the 3-year and 5-year windows.

The HDMA via *HIMA* did not identify any CpG sites as significant mediators. In the HDMA using a combination of *DACT* and causal mediation analysis, we identified twenty-two CpG sites to mediate the positive association between  $\text{PM}_{2.5}$  exposure and ABC score (Table 3), while none were observed for Braak stage and CERAD score. One CpG site (cg16342341, *SORBS2* gene) was associated with all three exposure windows (1, 3 and 5-years prior to death), and eight with two exposure windows. Of note, cg16342341 (*SORBS2*) was also identified in the

MITM approach described above. The total effect estimated for all mediation analyses was positive but insignificant in this subsample of the cohort (see Christensen et al. 2023 for the significant total effect in the full cohort). The statistics of all CpG sites detected by *DACT* are summarized in the Supplement (Table S2).

### *Secondary analyses*

A gene ontology analysis was conducted for the top 1000 CpG sites associated  $PM_{2.5}$  and for the top 1000 CpG sites associated with the neuropathology markers. None of the KEGG pathways reached significance after correcting for multiple tests. Therefore, we summarized the top 10 KEGG pathways for each of the  $PM_{2.5}$  exposures or neuropathology markers in the Supplement (Table S3). One pathway, which is the longevity regulating pathway, was associated with both 3-year exposure to  $PM_{2.5}$  and CERAD score. Eight genes (*HSPA1A*, *HSPA1L*, *IRS1*, *KRAS*, *NRAS*, *RPTOR*, *IRS2*, *ATG5*) in this pathway were enriched by differentially methylated CpG sites that were associated with 3-year  $PM_{2.5}$  exposure, and ten genes (*ADCY3*, *ADCY5*, *NFKB1*, *PRKAG2*, *RPTOR*, *TSC2*, *EHMT1*, *ULK1*, *AKT1S1*, *ATG5*) with CERAD score. Of note, *AKT1S1* was also among the genes that were identified in the HDMA (*DACT* and causal mediation analysis).

## Discussion

In the current study of 159 donors from the Emory Goizueta ADRC brain bank, we identified differential DNAm in prefrontal cortex tissues at two CpG sites to be significantly associated with long-term  $PM_{2.5}$  exposure. The two CpG sites [cg25433380 (intergenic) and cg10495669 (*RBCK1*)] that were associated with  $PM_{2.5}$  exposure were consistently associated with long-term exposures to traffic-related  $PM_{2.5}$  1 year, 3 years, and 5 years prior to death, after controlling for measured and unmeasured confounding. While cg25433380 and cg10495669 were not



associated with increases in neuropathology markers, we identified four CpG sites that overlapped between the top 1000 CpG sites associated with PM<sub>2.5</sub> and neuropathology markers (MITM approach) and 22 CpG sites that mediated the adverse effect of PM<sub>2.5</sub> exposures on AD-related neuropathology markers using HDMA. In addition, the longevity regulating pathway, was found to be enriched by differentially methylated CpG sites associated with PM<sub>2.5</sub> (3-year exposure window) and CERAD score.

Although there is a growing body of research on PM<sub>2.5</sub>-associated DNAm patterns in the human blood,<sup>71</sup> this is the first study showing an association between PM<sub>2.5</sub> exposure and differential DNAm in the brain (cg25433380 and cg10495669). Scarce evidence related to air pollution has been reported so far on cg25433380. On the other hand, higher DNA methylation levels of cg10495669 in nasal cells have been associated with 1-year ambient PM<sub>2.5</sub> exposure among 503 children from Project Viva in Massachusetts state.<sup>104</sup> *RBCK1*, the gene which cg10495669 is assigned to, is involved in carcinogenesis and inflammation pathways. The overexpression of *RBCK1* was observed in multiple cancer cells, including renal, colorectal, and breast cells, in *in-vitro* experiments.<sup>105-107</sup> The knockdown of *RBCK1* in renal cancer cells may induce p53 expressions, and thus, Yu et al. proposed a model in which *RBCK1* promoted the ubiquitination and degradation of p53, a protein playing a major role in DNA damage response.<sup>107</sup> The impairment of p53 expression and activity might participate in neurodegeneration, as p53 can bind to genes that regulate expression of synaptic proteins, neurite outgrowth, and axonal regeneration, which indicated a neuroprotective role against AD development.<sup>108</sup> In addition, *RBCK1*, as part of linear ubiquitin chain assembly complex, can regulate the proinflammatory-cytokines-induced nuclear factor kappa B (NF-κB) activation which serves as a pivotal mediator of inflammatory responses.<sup>109,110</sup> NF-κB activation is a common feature of many neurodegenerative diseases,<sup>111</sup> and the increased expression and/or activation of NF-κB has been largely observed in post-mortem studies of AD patients.<sup>112</sup> However, the two CpG sites were not found to be associated with any neuropathology markers in the current

analysis. More research is warranted on these CpG sites to investigate their potential role in AD development with a larger sample size and participants of more diverse disease stages from preclinical to severe dementia.

We identified four CpG sites (cg01835635, cg09830308, cg16342341, and cg27459981) that overlapped between the top 1000 CpG sites associated with both PM<sub>2.5</sub> and neuropathology markers via MITM approach. Three of these CpGs (cg16342341, cg09830308 and cg27459981) or their related genes have been previously associated with AD or PM<sub>2.5</sub> exposure. Cg09830308 and cg27459981, assigned to the mixed lineage kinase domain like pseudokinase gene (*MLKL*), were both associated with Braak stage and PM<sub>2.5</sub> exposure 3 and 5 years prior to death. *MLKL* plays a critical role in TNF-induced cell death (i.e., necroptosis). Caccamo et al. found that necroptosis was activated in postmortem brains of AD patients and positively correlated with Braak stage, and *MLKL* expression was significantly higher compared to control cases' brain tissues.<sup>113</sup> Similarly, Jayaraman et al. reported that necroptosis signaling was highly activated in the hippocampus of AD patients, as illustrated by the increased mRNA expression of genes, including *MLKL*, that encode key proteins involved in the execution of necroptosis.<sup>114</sup> Shigemizu et al. detected the genetic variant of *MLKL* associated with AD among 3777 Japanese subjects over 59 years old via genome-wide gene-based burden testing on rare coding variants, and they demonstrated that the loss-of-function variant of *MLKL* played a crucial role in AD pathogenesis via in-vitro experiments.<sup>115</sup> Furthermore, Wang et al. demonstrated that the knockdown of *MLKL* significantly increased the ratio of A $\beta$ 42 to A $\beta$ 40 in an AD model HEK293 cell line.<sup>116</sup> The ratio was used as a potential diagnostic marker of AD.<sup>117</sup> Collectively, traffic-related PM<sub>2.5</sub> exposure might induce the TNF-mediated neuroinflammation, resulting in necroptosis, and thus contribute to AD pathogenesis.

Cg16342341, assigned to the Sorbin and SH3 domain-containing protein 2 gene (*SORBS2*), was also identified as potential mediator in the HDMA, where it mediated the association of all PM<sub>2.5</sub> exposure windows with ABC score. As *SORBS2* is well known for its role in AD and

neuroinflammation<sup>118 119</sup> and has also been associated with PM<sub>2.5</sub> exposure in rats<sup>120</sup>, our findings contribute to the growing body of evidence of *SORBS2* expression playing a role in PM<sub>2.5</sub> associated changes in neuropathology markers of AD. *SORBS2* was found to repress IL-6 and TNF- $\alpha$  expression in the mouse embryonic fibroblasts,<sup>121</sup> and Chen et al. demonstrated that the level of *SORBS2* was lower in the brains of AD model mice compared to wild type mice,<sup>118</sup> implying a role of *SORBS2* in regulating neuroinflammation. In a human study of families multiply affected by AD, Lee et al. reported that genetic variation in *SORBS2* was associated with age at onset of AD.<sup>119</sup> While evidence on the association between PM<sub>2.5</sub> exposure and *SORBS2* is more scarce, Chao et al. reported that prenatal exposure to PM<sub>2.5</sub> caused upregulation of microRNAs targeting *SORBS2* gene in fetal rat cortex tissues<sup>120</sup>.

In addition to cg16342341, we identified 21 other CpGs as potential mediators of the association between long-term exposure to traffic-related PM<sub>2.5</sub> and ABC score using HDMA, and two of these CpGs have been previously reported in association with AD. Differential methylation in cg07963191, assigned to the dual 3',5'-cyclic-AMP and -GMP phosphodiesterase 11A gene (*PDE11A*), mediated the adverse effect of the average PM<sub>2.5</sub> exposure 3 years prior to death on the ABC score. *PDE11A* pertains to the phosphodiesterase family that plays an essential role in neuroplasticity and neuroprotection.<sup>122</sup> Differential methylation in cg27297993, assigned to the gamma-aminobutyric acid B receptor 1 gene (*GABBR1*), mediated the adverse effect of the average PM<sub>2.5</sub> exposure 3 and 5 years prior to death on the ABC score. *GABBR1* is the main inhibitory neurotransmitter in the central nervous system, which was reported to be downregulated in the brains of AD patients.<sup>123</sup> Iwakiri et al. observed a negative correlation between *GABBR1* and NFT formation in the hippocampus of 16 aged subjects, suggesting that the increased or stable expression of *GABBR1* may contribute to neuronal resistance to AD development.<sup>124</sup>

To derive more functional insights from the mediating CpG sites, we conducted gene ontology analysis based on KEGG pathway database for the top 1000 CpGs associated with

PM<sub>2.5</sub> exposure or neuropathology markers.<sup>99</sup> Proline-rich AKT1 substrate 1 (*AKT1S1*) was one of the genes enriched in the longevity regulating pathway which was found to overlap between PM<sub>2.5</sub> exposure and CERAD score. Of note, differential DNA methylation in cg00633834, which is assigned to *AKT1S1*, was also identified as potential mediator in the HDMA. *AKT1S1* can activate mammalian target of rapamycin (mTOR)–mediated signaling pathways when phosphorylated,<sup>125</sup> and mTOR signaling was observed to have higher activity in AD brains.<sup>126</sup> As mTOR plays a role for maintaining the balance between protein synthesis and degradation, Salvatore Oddo suggested a critical role of mTOR in the accumulation of A $\beta$  and tau proteins over the course of AD development from early to late stage.<sup>126</sup>

Our study has several strengths. We established for the first time a potential mediation effect of DNAm for the association between PM<sub>2.5</sub> and neuropathological changes of AD. Although false discovery is a problem in high-dimensional settings, we minimized the possibility of false discovery by verifying the indirect effect of CpG sites identified by HDMA using causal mediation analysis. The neuropathological changes of AD were quantified via multiple markers, including Braak stage, CERAD score, and ABC score, which covers the essential components (i.e., NFTs and A $\beta$  plaques) for the neuropathological diagnosis of AD. Further, the neuropathology markers were assessed by experienced neuropathologists at Emory Goizueta ADRC following a standardized protocol, which minimized the misclassification bias of outcomes. Finally, the high-resolution PM<sub>2.5</sub> exposure assessment model enabled the characterization of spatial variation in individual exposure and reduced the potential measurement error.<sup>33</sup>

Our study is not without limitations. First, the temporal sequence between mediators (DNAm changes) and outcomes (AD neuropathology) could not be clearly defined. Second, traffic-related PM<sub>2.5</sub> exposure was estimated based on the residential address of donors at death. Moving shortly prior to death could have introduced measurement errors in exposure assessment, and the selection of exposure windows was arbitrary, as the disease process of AD may start many years before death and vary by patients. Third, the results are from a single

brain bank and participants with a high APOE  $\epsilon 4$  carrier rate, so the generalizability should be tested in other brain bank or autopsy cohorts. Fourth, even though most of the study population was White, and we controlled for race, the ancestry effect on DNA methylation might persist as residual confounding. Finally, the current analysis only focused on the health effect of PM<sub>2.5</sub>, while other air pollutants such as nitrogen oxides or ozone might also play a role for AD.<sup>127,128</sup> Furthermore, PM<sub>2.5</sub> is a complex mixture and its composition varies by geographic region. From our analysis, we cannot determine which components, such as heavy metals, are driving the association with AD.<sup>129</sup>

## Conclusions

Using a combination of Meet-in-the-Middle approach, high-dimensional mediation analysis, and causal mediation analysis, we identified several CpG sites mediating the adverse effects of long-term exposure to traffic-related PM<sub>2.5</sub> exposure on the levels of AD-related neuropathology markers among prefrontal cortex tissues from 159 donors. Of note, several of these CpGs were identified by both approaches and located in genes related to neuroinflammation and neuroinflammation-mediated necroptosis. Our findings provide important information on the biological mechanisms underlying the PM<sub>2.5</sub> toxicity on AD pathogenesis. Future studies evaluating the mediating role of DNAm on AD-related outcomes should consider: 1) performing the analysis among early-stage AD patients or patients with mild cognition impairment to further illustrate the role of PM<sub>2.5</sub> in AD etiology; 2) performing genome-wide DNAm together with transcriptomics, proteomics, and/or metabolomics to capture a holistic picture of the underlying mechanism. Future work should investigate these findings in separate brain banks and determine whether the DNAm changes are found in other more accessible tissues with the goal of informing biomarkers of PM<sub>2.5</sub> exposure that may be relevant to AD development.

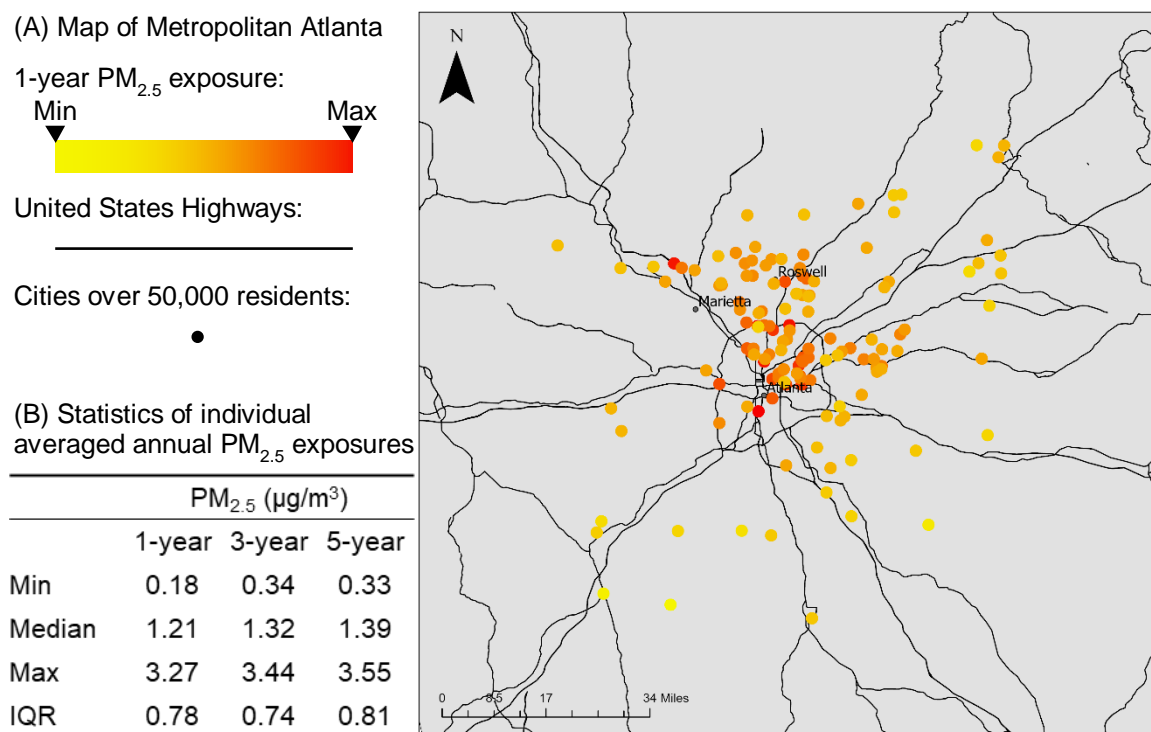
## Tables and figures

**Table 1.** Selected population characteristics among the donors included in the current analysis.

	N=159
Age at death, mean (SD)	76.6 (9.98)
Sex, No. (%)	
Female	70 (44.0)
Male	89 (56.0)
Race, No. (%)	
Black	17 (10.7)
White	142 (89.3)
Educational attainment, No. (%)	
High school or less	36 (22.6)
College degree	76 (47.8)
Graduate degree or more	47 (29.6)
Area Deprivation Index, mean (SD)	36.3 (24.2)
Diagnosis of dementia	
AD	86 (54.1)
Other dementia	66 (41.5)
No dementia	7 (4.4)
APOE genotype	
No $\epsilon$ 4 allele	70 (44.0)
Single $\epsilon$ 4 allele	68 (42.8)
Two $\epsilon$ 4 allele	21 (13.2)
Postmortem interval (hours), mean (SD)	11.7 (9.68)
Proportion of neuronal cells (%), mean (SD)	31.9 (8.21)
Braak stage, No. (%)	
Stage 1	16 (10.1)
Stage 2	11 (6.9)
Stage 3	20 (12.6)
Stage 4	17 (10.7)
Stage 5	22 (13.8)
Stage 6	73 (45.9)
CERAD score	
No	35 (22.0)
Sparse	4 (2.5)

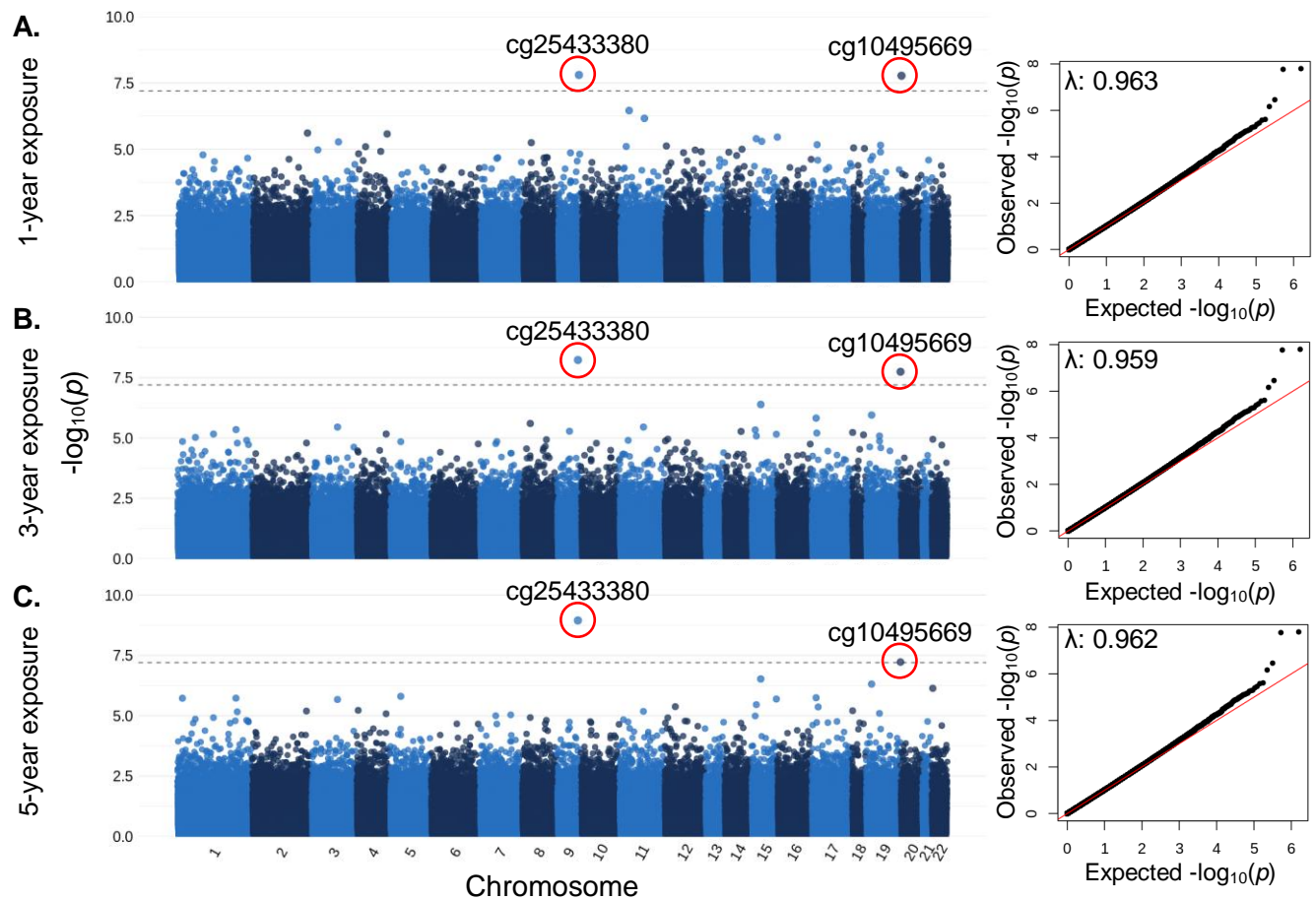
	N=159
Moderate	10 (6.3)
Frequent	110 (69.2)
ABC score	
Not	15 (9.4)
Low	29 (18.2)
Intermediate	22 (13.8)
High	93 (58.5)

Abbreviations: SD, standard deviation; AD, Alzheimer's disease; APOE, apolipoprotein E; CERAD, Consortium to Establish a Registry for AD; ABC, a combination of Amyloid, Braak stage, and CERAD (ABC) score.



**Figure 1.** Statistics and distribution of  $PM_{2.5}$  exposures in Metropolitan Atlanta (study area), Georgia, United States. (A) Map of Metropolitan Atlanta with individual 1-year averaged annual  $PM_{2.5}$  exposure. The dots denote the donors' residential address and are colored according to their  $PM_{2.5}$  exposures as showed in the legend. Red means a higher exposure level. (B) Statistics of individual averaged annual  $PM_{2.5}$  exposures for 1 year, 3 years, and 5 years.





**Figure 2.** Manhattan and QQ plots for the epigenome-wide association of PM<sub>2.5</sub> exposures (A. 1-year / B. 3-year / C. 5-year average exposure prior to death) and DNA methylation in postmortem frontal cortex tissue.  $\lambda$  denotes the inflation factor. Adjusted for covariates: age at death, sex, race, educational attainment, post-mortem interval, area deprivation index, and cell type composition. Unmeasured confounding and bias were adjusted with surrogate variable analysis and R package *Bacon*. Bonferroni threshold: 0.05/789,286.

**Table 2.** CpGs associated with traffic-related PM<sub>2.5</sub> exposure prior to death and their association with neuropathology markers.

CpG	chr	Position	Gene		Coefficients <sup>a</sup>	p-values <sup>b</sup>
A. CpGs with PM <sub>2.5</sub> exposures						
cg25433380	9	388,531	Intergenic	1-year exposure	0.0065	1.58×10 <sup>-8</sup>
				3-year exposure	0.0066	5.82×10 <sup>-9</sup>
				5-year exposure	0.0063	1.12×10 <sup>-9</sup>
cg10495669	20	137,531,767	RBCK1	1-year exposure	0.0127	1.69×10 <sup>-8</sup>
				3-year exposure	0.0128	1.78×10 <sup>-8</sup>
				5-year exposure	0.0114	5.96×10 <sup>-8</sup>
B. CpGs with neuropathology markers						
cg25433380	9	388,531	Intergenic	Braak stage	0.08	0.729
				CERAD	0.05	0.629
				ABC	0.04	0.825
cg10495669	20	137,531,767	RBCK1	Braak stage	0.02	0.929
				CERAD	0.12	0.397
				ABC	0.09	0.593

Abbreviations: PM<sub>2.5</sub>, fine particulate matter; chr, chromosome; RBCK1, RanBP-type and C3HC4-type zinc finger-containing protein 1.

<sup>a</sup> The coefficients for PM<sub>2.5</sub> exposures represent the change in the beta values of CpG sites associated with one-unit increase in the exposures; the coefficients for neuropathology markers represent the change in the neuropathology markers associated with one-interquartile-range increase in the beta values of CpG sites.

<sup>b</sup> The Bonferroni threshold:  $0.05/789,286 \approx 6.33 \times 10^{-8}$ .

**Table 3.** Indirect effect estimated by causal mediation analysis via the R package *mediation* of CpG sites selected by high-dimensional mediation analysis for the associations between PM<sub>2.5</sub> exposure and ABC score <sup>a</sup>.

CpG	chr	Gene	Exposure	DACT p-values <sup>b</sup>	ACME <sup>c</sup>	Total effect <sup>d</sup>
cg23932332	1	DUSP10	3-year	4.36E-08	0.056 (0.005, 1.50E-01)	0.086 (-0.110, 2.80E-01)
			5-year	2.98E-08	0.060 (0.002, 1.70E-01)	0.104 (-0.081, 3.10E-01)
cg08512806	1	TARBP1	3-year	5.30E-08	0.058 (0.008, 1.30E-01)	0.084 (-0.107, 3.00E-01)
			5-year	3.82E-08	0.063 (0.009, 1.30E-01)	0.102 (-0.080, 3.10E-01)
cg10705045	2	RNF144A	5-year	2.63E-08	0.063 (0.001, 1.40E-01)	0.109 (-0.079, 3.10E-01)
cg17275287	2	Intergenic	3-year	3.43E-09	0.085 (0.019, 1.70E-01)	0.079 (-0.118, 3.00E-01)
			5-year	2.00E-09	0.089 (0.020, 1.80E-01)	0.097 (-0.093, 3.00E-01)
cg07258300	2	CYP27C1	3-year	5.43E-08	0.080 (0.020, 1.50E-01)	0.083 (-0.107, 3.00E-01)
cg05532414	2	Intergenic	3-year	6.25E-08	0.071 (0.004, 1.70E-01)	0.090 (-0.084, 3.20E-01)
cg07963191	2	PDE11A	3-year	3.10E-08	0.061 (0.005, 1.40E-01)	0.080 (-0.103, 3.00E-01)
cg26109897	4	TBC1D14	3-year	2.13E-08	0.085 (0.010, 1.90E-01)	0.090 (-0.098, 3.10E-01)
cg26877022	4	POLR2B	3-year	4.54E-09	0.080 (0.015, 1.80E-01)	0.089 (-0.092, 3.10E-01)
			5-year	1.26E-08	0.077 (0.011, 1.80E-01)	0.107 (-0.079, 3.10E-01)
cg16342341	4	SORBS2	1-year	1.35E-09	0.097 (0.021, 1.80E-01)	0.034 (-0.168, 2.30E-01)
			3-year	5.45E-09	0.076 (0.017, 1.60E-01)	0.080 (-0.106, 2.80E-01)
			5-year	1.61E-09	0.078 (0.017, 1.50E-01)	0.098 (-0.093, 3.20E-01)
cg17444747	5	COL23A1	5-year	3.26E-08	0.074 (0.015, 1.50E-01)	0.098 (-0.085, 2.90E-01)
cg27297993	6	GABBR1	3-year	8.30E-09	0.064 (0.009, 1.40E-01)	0.084 (-0.091, 3.00E-01)
			5-year	9.26E-09	0.066 (0.003, 1.40E-01)	0.103 (-0.076, 3.00E-01)
cg00829961	8	Intergenic	3-year	1.37E-08	0.075 (0.009, 1.70E-01)	0.092 (-0.092, 3.10E-01)
			5-year	3.21E-08	0.075 (0.012, 1.70E-01)	0.110 (-0.078, 3.30E-01)
cg02987635	10	C10orf11	3-year	4.14E-08	0.063 (0.004, 1.50E-01)	0.079 (-0.099, 3.00E-01)
cg06805557	11	APBB1	5-year	4.14E-08	0.062 (0.007, 1.30E-01)	0.101 (-0.104, 3.00E-01)
cg19969778	11	SIAE; SPA17	3-year	8.97E-09	0.065 (0.008, 1.30E-01)	0.080 (-0.108, 3.10E-01)
			5-year	1.85E-08	0.063 (0.010, 1.30E-01)	0.098 (-0.092, 3.10E-01)
cg20713102	15	ZSCAN2	5-year	5.05E-08	0.074 (0.014, 1.60E-01)	0.106 (-0.083, 3.10E-01)
cg09088153	15	Intergenic	3-year	4.69E-08	0.072 (0.013, 1.50E-01)	0.089 (-0.094, 3.20E-01)
cg27181554	16	SEPX1	1-year	1.79E-08	0.084 (0.021, 1.80E-01)	0.039 (-0.162, 2.70E-01)
			3-year	2.73E-08	0.069 (0.015, 1.50E-01)	0.085 (-0.108, 2.80E-01)
cg20389589	16	FAM57B	3-year	2.97E-08	0.069 (0.003, 1.60E-01)	0.084 (-0.120, 2.90E-01)
cg06832209	16	ADGRG3	3-year	4.26E-08	0.078 (0.015, 1.60E-01)	0.089 (-0.101, 2.80E-01)
cg00633834	19	AKT1S1; TBC1D17	5-year	3.72E-08	0.081 (0.017, 1.60E-01)	0.095 (-0.090, 2.90E-01)

Abbreviations: PM<sub>2.5</sub>, fine particulate matter; chr, chromosome; ACME, average causal

mediated effect (i.e., indirect effect).

<sup>a</sup> All CpG sites that were selected by *DACT* and had a positive ACME were associated with ABC score.

<sup>b</sup> The *p*-values of mediation effect testing conducted by *DACT*.

<sup>c</sup> The ACME was associated with one-interquartile-range increase in beta values of CpG sites.

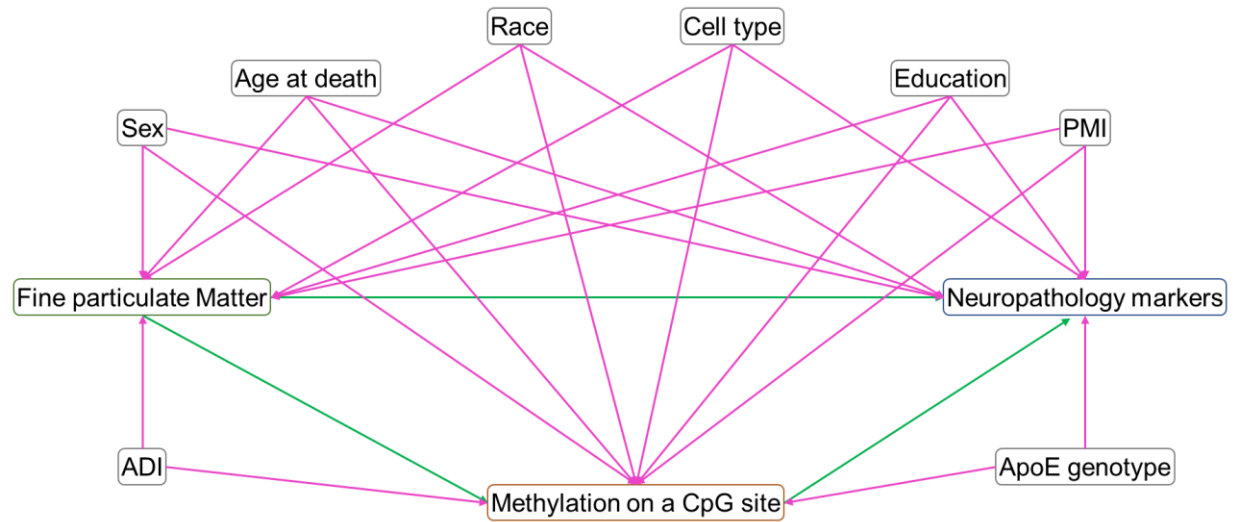
<sup>d</sup> Effect estimates, associated with 1-unit increase, of PM<sub>2.5</sub> exposures on neuropathology markers.

## Appendix for Chapter 3

### Methods

#### *Genome-wide DNA methylation quality control and processing*

All DNAm data were preprocessed to identify mislabeled and low-quality samples, exclude specific probes, and reduce the impact of batch effects. Raw intensity files were converted to methylation beta values ranging on a continuous scale from 0 to 1 for each of the CpG sites measured on the array. The Illumina's 636 control probes were used via the R package *ewastools* to assess technique parameters including array staining, extension, hybridization, target removal, specificity, and bisulfite conversion.<sup>130</sup> Additional sample outlier detection was implemented based on detection *p* value, beadcount, and distance from the group average in principal components. All samples passed the aforementioned checks. After this, participant sex and replicate status were also verified using XY probes and SNP probes, respectively. Then, the Funnorm function and Combat function was used to normalize the distributions to reduce technical variation and correct for differences between type I and type II probe signals. The following probes were further removed: XY probes, low-quality probes with missing in more than 5% of samples, probes with poor detection *p*-values, probes predicted to cross-hybridize, probes that bind to the sex chromosomes, polymorphic probes, and probes with infinite values. In total, after all preprocessing steps, 159 samples and 789,286 CpG sites remained for the down-stream analysis. We used estimateCellCounts function in the R package *minfi* to obtain the cell-type proportions (neuronal vs. non-neuronal cells) using the reference dataset published previously.<sup>86,87</sup>



**Figure S1.** Directed acyclic graph of the confounding structure for the association between fine particulate matter exposure and neuropathology markers. PMI, post-mortem interval; ADI, area deprivation index; APOE, Apolipoprotein E.

**Table S1.** The overlapping CpGs associated with both traffic-related PM<sub>2.5</sub> exposure prior to death and neuropathology markers.

CpG	chr	Position	Gene		Coefficients <sup>a</sup>	p-values
A. CpGs with PM <sub>2.5</sub> exposures						
cg16342341	4	186,571,988	SORBS2	1-year exposure	0.0102	1.27×10 <sup>-3</sup>
cg01835635	11	116,693,535	APOA4	1-year exposure	-0.0168	6.16×10 <sup>-4</sup>
				3-year exposure	-0.0155	1.29×10 <sup>-3</sup>
cg09830308	16	74,734,321	MLKL	1-year exposure	0.0163	1.15×10 <sup>-3</sup>
				3-year exposure	0.0161	9.44×10 <sup>-4</sup>
				5-year exposure	0.0151	8.47×10 <sup>-4</sup>
cg27459981	16	74,734,571	MLKL	3-year exposure	0.0155	1.11×10 <sup>-3</sup>
				5-year exposure	0.0142	1.37×10 <sup>-3</sup>
B. CpGs with neuropathology markers						
cg16342341	4	186,571,988	SORBS2	CERAD	0.52	6.03×10 <sup>-4</sup>
				ABC	0.44	5.37×10 <sup>-4</sup>
cg01835635	11	116,693,535	APOA4	CERAD	-0.63	4.36×10 <sup>-4</sup>
cg09830308	16	74,734,321	MLKL	Braak stage	-0.51	3.09×10 <sup>-4</sup>
cg27459981	16	74,734,571	MLKL	Braak stage	-0.82	1.70×10 <sup>-5</sup>

Abbreviations: PM<sub>2.5</sub>, fine particulate matter; chr, chromosome; *SORBS2*, sorbin and SH3

domain-containing protein 2; *APOA4*, Apolipoprotein A-IV; *MLKL*, mixed lineage kinase domain like pseudokinase.

<sup>a</sup> The coefficients for PM<sub>2.5</sub> exposures represent the change in the beta values of CpG sites associated with one-unit increase in the exposures; the coefficients for neuropathology markers represent the change in the neuropathology markers associated with one-interquartile-range increase in the beta values of CpG sites.

**Table S2.** Indirect effect estimated by causal mediation analysis via the R package *mediation* of CpG sites selected by high-dimensional mediation analysis for the associations of PM<sub>2.5</sub> exposure with Braak stage and ABC score <sup>a</sup>.

CpG	chr	Gene	Exposure	Outcome	ACME <sup>c</sup>	Total effect <sup>d</sup>
cg09579061	1	<i>ADORA1</i>	1-year	Braak	-0.183 (-0.348, -5.00E-02)	0.032 (-0.318, 4.30E-01)
cg23932332	1	<i>DUSP10</i>	3-year	ABC	0.056 (0.005, 1.50E-01)	0.086 (-0.110, 2.80E-01)
			5-year	ABC	0.060 (0.002, 1.70E-01)	0.104 (-0.081, 3.10E-01)
cg08512806	1	<i>TARBP1</i>	3-year	ABC	0.058 (0.008, 1.30E-01)	0.084 (-0.107, 3.00E-01)
			5-year	ABC	0.063 (0.009, 1.30E-01)	0.102 (-0.080, 3.10E-01)
cg10705045	2	<i>RNF144A</i>	5-year	ABC	0.063 (0.001, 1.40E-01)	0.109 (-0.079, 3.10E-01)
cg17275287	2	Intergenic	3-year	ABC	0.085 (0.019, 1.70E-01)	0.079 (-0.118, 3.00E-01)
			5-year	ABC	0.089 (0.020, 1.80E-01)	0.097 (-0.093, 3.00E-01)
cg05532414	2	Intergenic	3-year	ABC	0.071 (0.004, 1.70E-01)	0.090 (-0.084, 3.20E-01)
cg07963191	2	<i>PDE11A</i>	3-year	ABC	0.061 (0.005, 1.40E-01)	0.080 (-0.103, 3.00E-01)
cg26109897	4	<i>TBC1D14</i>	3-year	ABC	0.085 (0.010, 1.90E-01)	0.090 (-0.098, 3.10E-01)
cg26877022	4	<i>POLR2B</i>	3-year	ABC	0.080 (0.015, 1.80E-01)	0.089 (-0.092, 3.10E-01)
			5-year	ABC	0.077 (0.011, 1.80E-01)	0.107 (-0.079, 3.10E-01)
			1-year	ABC	0.097 (0.021, 1.80E-01)	0.034 (-0.168, 2.30E-01)
cg16342341	4	<i>SORBS2</i>	3-year	ABC	0.076 (0.017, 1.60E-01)	0.080 (-0.106, 2.80E-01)
			5-year	ABC	0.078 (0.017, 1.50E-01)	0.098 (-0.093, 3.20E-01)
cg17444747	5	<i>COL23A1</i>	5-year	ABC	0.074 (0.015, 1.50E-01)	0.098 (-0.085, 2.90E-01)
cg27297993	6	<i>GABBR1</i>	3-year	ABC	0.064 (0.009, 1.40E-01)	0.084 (-0.091, 3.00E-01)
			5-year	ABC	0.066 (0.003, 1.40E-01)	0.103 (-0.076, 3.00E-01)
cg00829961	8	Intergenic	3-year	ABC	0.075 (0.009, 1.70E-01)	0.092 (-0.092, 3.10E-01)
			5-year	ABC	0.075 (0.012, 1.70E-01)	0.110 (-0.078, 3.30E-01)
cg02987635	10	<i>C10orf11</i>	3-year	ABC	0.063 (0.004, 1.50E-01)	0.079 (-0.099, 3.00E-01)
cg06805557	11	<i>APBB1</i>	5-year	ABC	0.062 (0.007, 1.30E-01)	0.101 (-0.104, 3.00E-01)
cg19969778	11	<i>SIAE;</i> <i>SPA17</i>	3-year	ABC	0.065 (0.008, 1.30E-01)	0.080 (-0.108, 3.10E-01)
			5-year	ABC	0.063 (0.010, 1.30E-01)	0.098 (-0.092, 3.10E-01)
cg17562250	14	<i>LOC101927124;</i> <i>HEATR5A</i>	1-year	ABC	-0.077 (-0.174, -1.00E-02)	0.039 (-0.169, 2.60E-01)
			3-year	ABC	-0.065 (-0.169, 0.00E+00)	0.085 (-0.099, 3.00E-01)
cg20713102	15	<i>ZSCAN2</i>	5-year	ABC	0.074 (0.014, 1.60E-01)	0.106 (-0.083, 3.10E-01)
cg09088153	15	Intergenic	3-year	ABC	0.072 (0.013, 1.50E-01)	0.089 (-0.094, 3.20E-01)
cg27181554	16	<i>SEPX1</i>	1-year	ABC	0.084 (0.021, 1.80E-01)	0.039 (-0.162, 2.70E-01)
			3-year	ABC	0.069 (0.015, 1.50E-01)	0.085 (-0.108, 2.80E-01)
cg20389589	16	<i>FAM57B</i>	3-year	ABC	0.069 (0.003, 1.60E-01)	0.084 (-0.120, 2.90E-01)
cg06832209	16	<i>ADGRG3</i>	3-year	ABC	0.078 (0.015, 1.60E-01)	0.089 (-0.101, 2.80E-01)
			1-year	ABC	-0.072 (-0.161, -1.00E-02)	0.030 (-0.171, 2.40E-01)
cg09830308	16	<i>MLKL</i>	1-year	Braak	-0.152 (-0.319, -1.00E-02)	0.023 (-0.318, 3.80E-01)
			3-year	ABC	-0.071 (-0.151, 0.00E+00)	0.073 (-0.130, 2.90E-01)
			5-year	ABC	-0.073 (-0.163, -1.00E-02)	0.092 (-0.115, 3.10E-01)
cg27459981	16	<i>MLKL</i>	1-year	ABC	-0.089 (-0.178, -2.00E-02)	0.028 (-0.198, 2.50E-01)



			1-year	Braak	-0.182 (-0.364, -4.00E-02)	0.019 (-0.333, 3.60E-01)
			3-year	ABC	-0.086 (-0.168, -2.00E-02)	0.071 (-0.132, 2.80E-01)
			5-year	ABC	-0.087 (-0.179, -2.00E-02)	0.089 (-0.093, 2.90E-01)
cg08591058	17	Intergenic	1-year	ABC	-0.092 (-0.175, -2.00E-02)	0.030 (-0.162, 2.50E-01)
			3-year	ABC	-0.084 (-0.165, -2.00E-02)	0.073 (-0.125, 2.90E-01)
cg26535871	17	<i>MGAT5B</i>	3-year	ABC	-0.063 (-0.147, 0.00E+00)	0.070 (-0.129, 2.80E-01)
			5-year	ABC	-0.065 (-0.164, -1.00E-02)	0.089 (-0.114, 2.90E-01)
cg00633834	19	<i>AKT1S1;</i> <i>TBC1D17</i>	5-year	ABC	0.081 (0.017, 1.60E-01)	0.095 (-0.090, 2.90E-01)

---

Abbreviations: PM<sub>2.5</sub>, fine particulate matter; chr, chromosome; ACME, average causal

mediated effect (i.e., indirect effect).

<sup>a</sup> All CpG sites that were selected by *DACT* and had a significant ACME were associated with Braak stage or ABC score.

<sup>b</sup> The *p*-values of mediation effect testing conducted by *DACT*.

<sup>c</sup> The ACME was associated with one-interquartile-range increase in beta values of CpG sites.

<sup>d</sup> Effect estimates, associated with 1-unit increase, of PM<sub>2.5</sub> exposures on neuropathology markers.

**Table S3.** The top 10 KEGG pathways enriched by differentially methylated CpG sites that were associated with PM2.5 exposure and neuropathology markers, respectively.

Exposure/ outcome	KEGG terms	Description	N of genes in the KEGG term	N of differentially methylated genes	p-value	False discovery rate	Significant genes
1-year exposure	path:hsa04137	Mitophagy - animal	69	8	0.0106011	1	CSNK2B, MRAS, KRAS, NRAS, PRKN, AMBRA1, MAPK9, ATG5, USP15
	path:hsa00030	Pentose phosphate pathway	25	4	0.0145971	1	GPI, PGD, TALDO1, H6PD
	path:hsa04935	Growth hormone synthesis, secretion and action	117	13	0.0168878	1	GHR, GNAI2, IRS1, KRAS, MAP3K1, NRAS, PRKCB, MAPK9, MAP2K1, SSTR3, CACNA1C, IRS2, BCAR1
	path:hsa04140	Autophagy - animal	134	13	0.0175384	1	EIF2S1, ERN1, MRAS, IRS1, KRAS, NRAS, SH3GLB1, ATG16L1, AMBRA1, MAPK9, MAP2K1, RPTOR, IRS2, ATG5
	path:hsa05010	Alzheimer disease	359	23	0.0278593	1	APC2, COX7C, CSNK2B, DVL3, EIF2S1, ERN1, ATF6, GRIN2B, APP, IRS1, KRAS, LRP1, LRP5, NRAS, ATP5PO, AMBRA1, MAPK9, MAP2K1, PSMD13, WNT6, CACNA1C, AXIN1, IRS2, SLC39A13, COX7A2L
	path:hsa04720	Long-term potentiation	64	8	0.0319516	1	GRIA1, GRIN2B, KRAS, NRAS, PRKCB, MAP2K1, RAP1B, CACNA1C
	path:hsa04144	Endocytosis	244	19	0.0365904	1	EHD1, AGAP2, AP2A2, PSD3, PIP5K1C, LDLRAP1, CHMP4A, HSPA1A, HSPA1L, SH3GLB1, RUFY2, VPS35, PARD3, RAB5A, SH3GL1, SH3GL2, EEA1, PARD6G, RAB11A, ASAP2, IQSEC1
	path:hsa01200	Carbon metabolism	105	8	0.0463284	1	SDSL, PHGDH, GPI, PCCA, PGD, TALDO1, SUCLG2, H6PD
	path:hsa04213	Longevity regulating pathway - multiple species	61	7	0.0484809	1	HSPA1A, HSPA1L, IRS1, KRAS, NRAS, RPTOR, IRS2, ATG5
	path:hsa05034	Alcoholism	165	11	0.0489502	1	H4C16, GNAI2, GRIN2B, KRAS, NRAS, MAP2K1, H2AC14, H2AC16, H2BC14, H3C10, H4C3, HDAC4

3-year exposure	path:hsa04930	Type II diabetes mellitus	45	8	0.0045222	0.7312499	HK1, IRS1, PRKCZ, MAPK9, SLC2A4, CACNA1B, CACNA1C, IRS2
	path:hsa04213	Longevity regulating pathway - multiple species	61	9	0.0052512	0.7312499	ADCY1, HSPA1A, HSPA1L, IRS1, KRAS, NRAS, PRKAG2, RPTOR, IRS2, ATG5
	path:hsa00900	Terpenoid backbone biosynthesis	23	4	0.0062322	0.7312499	ZMPSTE24, FNTA, ICMT, IDI1
	path:hsa04720	Long-term potentiation	64	9	0.0103579	0.9114961	ADCY1, GRIA1, GRIN2B, KRAS, NRAS, PRKCB, MAP2K1, RAP1B, CACNA1C
	path:hsa04650	Natural killer cell mediated cytotoxicity	119	10	0.0180657	1	HCST, PTK2B, FYN, KRAS, NFATC2, NRAS, PAK1, PRKCB, MAP2K1, VAV2
	path:hsa04211	Longevity regulating pathway	88	10	0.0235191	1	ADCY1, EHMT2, CREB1, SESN3, IRS1, KRAS, NRAS, PRKAG2, RPTOR, IRS2, ATG5
	path:hsa05010	Alzheimer disease	359	23	0.0287495	1	APC2, COX7C, DVL3, EIF2S1, ATF6, GRIN2B, APP, IRS1, KRAS, LRP1, LRP5, ATP2A2, NRAS, ATP5PO, AMBRA1, MAPK9, MAP2K1, PSMA2, PSMD13, SLC39A8, WNT6, CACNA1C, AXIN1, CASP8, IRS2
	path:hsa04910	Insulin signaling pathway	131	12	0.0312015	1	HK1, ACACA, IRS1, KRAS, NRAS, PRKAG2, PRKCZ, MAPK9, MAP2K1, RPTOR, SLC2A4, IRS2
	path:hsa04144	Endocytosis	244	19	0.032739	1	EHD1, AP2A2, PSD3, PIP5K1C, CYTH4, CHMP4A, HSPA1A, HSPA1L, SH3GLB1, VTA1, RUFY2, PRKCZ, PARD3, RAB5A, WIPF3, SH3GL1, EEA1, RAB11A, DNAJC6, IQSEC1
	path:hsa04935	Growth hormone synthesis, secretion and action	117	12	0.0330168	1	ADCY1, CREB1, IRS1, KRAS, MAP3K1, NRAS, PRKCB, MAPK9, MAP2K1, SSTR3, CACNA1C, IRS2
5-year exposure	path:hsa00590	Arachidonic acid metabolism	61	7	0.0030161	1	PLB1, ALOX5, GGT1, PLA2G4D, GPX4, ALOXE3, PLA2G6
	path:hsa04720	Long-term potentiation	64	9	0.0134991	1	ADCY1, GRIA1, GRIN2B, NRAS, PPP3CA, PRKCB, MAP2K1, RAP1B, CACNA1C
	path:hsa04930	Type II diabetes mellitus	45	7	0.0195656	1	HK1, PRKCZ, MAPK9, SLC2A4, CACNA1B, CACNA1C, IRS2
	path:hsa04931	Insulin resistance	105	11	0.021487	1	PPARGC1B, CREB1, SLC27A1, NOS3, PRKAG2, PRKCB, PRKCZ, MAPK9, SLC2A4, IRS2, CREB5

	path:hsa00450	Selenocompound metabolism	17	3	0.0343345	1	TXNRD2, PSTK, PAPSS1
	path:hsa00513	Various types of N-glycan biosynthesis	38	5	0.0400335	1	HEXB, STT3A, ALG11, MAN1C1, CHST8, TUSC3
	path:hsa04370	VEGF signaling pathway	59	7	0.0412512	1	PLA2G4D, NFATC2, NOS3, NRAS, PPP3CA, PRKCB, MAP2K1
	path:hsa00900	Terpenoid backbone biosynthesis	23	3	0.0437082	1	ZMPSTE24, FNTA, ICMT
	path:hsa04136	Autophagy - other	30	4	0.0537717	1	WIPI2, ATG16L1, RPTOR, ATG5
	path:hsa00591	Linoleic acid metabolism	30	3	0.0537983	1	PLB1, PLA2G4D, PLA2G6
Braak stage	path:hsa05032	Morphine addiction	88	13	0.0011824	0.4161928	ADCY1, PDE10A, ADCY5, ADCY9, ADCY4, GNG7, GNGT2, GRK5, ARRB1, PDE11A, PDE1A, PDE3B, PDE4D, PRKCA
	path:hsa05110	Vibrio cholerae infection	49	7	0.0038668	0.6710898	ADCY9, ARF1, KCNQ1, ATP6V0C, PLCG2, PRKCA, TJP1, ATP6V0D1
	path:hsa04261	Adrenergic signaling in cardiomyocytes	148	15	0.0074374	0.6710898	CACNG3, ADCY1, ADCY5, ADCY9, CREB1, CREM, ADCY4, KCNQ1, PPP1CA, PPP2R1A, PRKCA, TPM3, CACNA1C, CACNB3, CACNA2D2
	path:hsa05130	Pathogenic Escherichia coli infection	189	15	0.0081416	0.6710898	WIPF2, TUBB, CYFIP1, CYTH4, IKBKB, ARF1, MYO1C, MYO5C, BAIAP2L1, RAC1, WIPF3, TJP1, TNFRSF1A, MYH14, TNFRSF10A
	path:hsa04015	Rap1 signaling pathway	207	19	0.0095325	0.6710898	ADCY1, ADCY5, ADCY9, CSF1R, CTNND1, ADCY4, EPHA2, FGF6, FLT4, FYB1, PRKD2, RAPGEF1, NGFR, EVL, PRKCA, PARD3, RAC1, VEGFC, FGF23, FGF18
	path:hsa04061	Viral protein interaction with cytokine and cytokine receptor	97	6	0.0146182	0.8213017	CSF1R, CX3CR1, IL6R, TNFRSF1A, TNFRSF10A, IL18R1
	path:hsa00230	Purine metabolism	123	11	0.0163327	0.8213017	ADCY1, PDE10A, ADCY5, ADCY9, ADCY4, AK4, AMPD3, GUCY2D, PDE11A, PDE1A, PDE3B, PDE4D
	path:hsa04151	PI3K-Akt signaling pathway	339	24	0.0215544	0.849279	PIK3AP1, COL1A1, CREB1, CSF1R, EPHA2, FGF6, PHLPP2, FLT4, GNG7, GNGT2, IKBKB, IL6R, LAMA3, MDM2, NGFR, NOS3, PPP2R1A, PRKCA, RAC1, CCND1, TNXB, VEGFC, FGF23, FGF18, CCND3

	path:hsa04010	MAPK signaling pathway	283	22	0.0217145	0.849279	CACNG3, CSF1R, DUSP4, EPHA2, FGF6, FLT4, IKBKB, ARRB1, NGFR, PRKCA, RAC1, RASGRF1, RPS6KA2, TNFRSF1A, VEGFC, CACNA1C, CACNB3, FGF23, MKNK1, FGF18, CACNA2D2, CD14
	path:hsa04014	Ras signaling pathway	227	18	0.0255507	0.8993838	CSF1R, EPHA2, FGF6, RASA3, RGL1, FLT4, GNG7, GNGT2, IKBKB, NGFR, PLCG2, PRKCA, RGL2, RAC1, RASGRF1, VEGFC, FGF23, FGF18, KSR1
<hr/>							
CERAD	path:hsa05130	Pathogenic Escherichia coli infection	189	18	0.0009299	0.3273211	TAB2, CYFIP1, GAPDH, CYFIP2, IL6, ARF1, MYO1A, MYO1C, MYO5A, MYO5B, NFKB1, CLDN18, BAIAP2L1, MAPK10, RAC1, TNFRSF1A, TRAF2, TUBB6
	path:hsa01523	Antifolate resistance	29	5	0.00559	0.8349026	ABCC4, IL6, NFKB1, SHMT1, TYMS, ABCG2
	path:hsa05146	Amoebiasis	98	10	0.0131476	0.8349026	LAMB4, LAMA1, IL6, ITGAM, NFKB1, NOS2, PRKCA, PTK2, ACTN4, ACTN1
	path:hsa05131	Shigellosis	238	18	0.0143556	0.8349026	RBCK1, TAB2, ARF1, MDM2, NFKB1, WIPI1, MAPK10, PTK2, RPTOR, RAC1, ELMO2, TNFRSF1A, TRAF2, ACTN4, TLN2, AKT1S1, ACTN1, ATG5
	path:hsa00430	Taurine and hypotaurine metabolism	16	3	0.0149302	0.8349026	FMO1, GGT7, BAAT
	path:hsa05171	Coronavirus disease - COVID-19	215	13	0.0162667	0.8349026	ADAR, RPL22L1, TAB2, RPL36, IFNAR1, IL6, NFKB1, PRKCA, MAPK10, RPL18, RPL26, STAT2, TNFRSF1A, TRAF3
	path:hsa04622	RIG-I-like receptor signaling pathway	65	6	0.0166032	0.8349026	ADAR, NFKB1, MAPK10, TRAF2, TRAF3, ATG5
	path:hsa03060	Protein export	23	3	0.0343889	1	SEC61A1, SEC61A2, SEC11C
	path:hsa05132	Salmonella infection	240	16	0.0352334	1	ACTR1B, TAB2, CYFIP1, PLEKHM2, GAPDH, CYFIP2, IL6, ARF1, NFKB1, DCTN4, MAPK10, RAC1, ELMO2, TNFRSF1A, TRAF2, TUBB6
	path:hsa04211	Longevity regulating pathway	88	9	0.0370466	1	ADCY3, ADCY5, NFKB1, PRKAG2, RPTOR, TSC2, EHMT1, ULK1, AKT1S1, ATG5
<hr/>							
ABC	path:hsa04672	Intestinal immune network for IgA production	44	5	0.0102832	1	HLA-DOA, IL6, IL15, CD28, CD40
	path:hsa04940	Type I diabetes mellitus	41	5	0.0141776	1	HLA-B, HLA-DOA, PTPRN, PTPRN2, CD28
	path:hsa05330	Allograft rejection	34	4	0.018464	1	HLA-B, HLA-DOA, CD28, CD40

path:hsa05332	Graft-versus-host disease	38	4	0.0199458	1	HLA-B, HLA-DOA, IL6, CD28
path:hsa00740	Riboflavin metabolism	8	2	0.0207243	1	ENPP1, ACP2
path:hsa05320	Autoimmune thyroid disease	46	4	0.0371603	1	HLA-B, HLA-DOA, CD28, CD40
path:hsa00130	Ubiquinone and other terpenoid-quinone biosynthesis	11	2	0.0407866	1	HPD, COQ5
path:hsa05131	Shigellosis	238	16	0.0428873	1	RBCK1, TAB2, MTOR, ITPR1, ARF1, MDM2, NFKB1, PLCG2, PTK2, RPTOR, RAC1, TNFRSF1A, C3, ACTN4, AKT1S1, ACTN1
path:hsa05150	Staphylococcus aureus infection	90	5	0.0528039	1	CFH, HLA-DOA, KRT9, SELPLG, C3
path:hsa04330	Notch signaling pathway	58	6	0.0547066	1	DTX3L, DTX1, HEYL, NOTCH2, ADAM17, TLE2, NCOR2

---

## Chapter 4 Metabolomic Signatures as Potential Mediators in the Relationship between Ambient Fine Particulate Matter and Pre- and Early-term Births in the Atlanta African American Maternal-Child Cohort

### Introduction

Ambient fine particulate matter (PM<sub>2.5</sub>) is a significant contributor to public health burden with adverse health effects well-characterized in all age groups.<sup>131-133</sup> Specifically, pregnant people and fetuses are more vulnerable to PM<sub>2.5</sub> exposure compared to the general population.<sup>10</sup> Existing evidence shows prenatal exposure to PM<sub>2.5</sub> to be associated with a series of adverse birth outcomes, including preterm (PTB) and early-term birth (ETB),<sup>10</sup> defined as being born prior to 37 weeks and 37-39 weeks' gestation, respectively.<sup>134</sup> PTB and ETB are among the leading contributors to neonatal morbidity and mortality,<sup>135</sup> and are associated with both short-term and long-term risks to child health.<sup>136</sup> Globally, approximately 10% of PTB cases are estimated to be attributable to ambient PM<sub>2.5</sub> exposure, with the highest burden in sub-Saharan Africa.<sup>137</sup>

Communities of color and low income communities in the United States (U.S.), especially among African Americans (AA), experience disproportionately higher rates of PTB and ETB, highlighting that health disparities begin *in utero*.<sup>138,139</sup> Although a significant body of research has explored the association between prenatal PM<sub>2.5</sub> exposure and PTB, reported findings have been limited among AA communities.<sup>140</sup> Some of the uncertainty in the results is due to the use of different exposure windows before and during pregnancy and the complex biological mechanisms involved in the etiology of PTB and ETB. Understanding the mechanisms underlying the toxicity of PM<sub>2.5</sub> exposure on the risk of PTB and ETB is essential to guiding policies and interventions to reduce the incidence of PTB and ETB among vulnerable populations.

High-resolution metabolomics has emerged as a powerful analytic platform in environmental health research used to characterize biological perturbations in the human metabolome associated with long-term and short-term exposures to air pollution.<sup>17</sup> A previous study found alterations of the mid-pregnancy serum metabolome in several oxidative stress and inflammation related pathways associated with air pollution exposure among 160 American mothers of multiple races.<sup>141</sup> In addition, researchers have employed metabolomics to identify biomarkers and pathways predictive of PTB, many of which were associated with PM<sub>2.5</sub> exposures.<sup>142-144</sup> These initial findings suggest that metabolomic changes may play a mediating role linking PM<sub>2.5</sub> with PTB and ETB, where prenatal PM<sub>2.5</sub> exposure may lead to an increased risk of PTB and ETB through changes in various intermediate metabolites, such as pro-inflammatory factors. By performing a mediation analysis, researchers can determine which metabolites mediate, and to what extent, the relationship of PM<sub>2.5</sub> exposure with PTB and ETB. The identification of metabolomic mediators in observational studies can strengthen the causal relationship between PM<sub>2.5</sub> exposure and PTB and ETB and, moreover, reveal the underlying biological mechanisms in a comprehensive perspective.<sup>27</sup> However, only a few epidemiological studies have employed metabolomics as intermediate variables or mediators investigating the health impacts of ambient air pollution on birth outcomes,<sup>145,146</sup> and none focus on AAs.

To address these critical knowledge gaps, we conducted this study in the Atlanta African American (ATL AA) Maternal-Child Cohort.<sup>147</sup> We used advanced high-resolution metabolomics and mediation analyses to identify metabolic perturbations (i.e., altered metabolites and biological pathways) that mediate the association of exposures to ambient PM<sub>2.5</sub> for four time windows (1-year prior to conception, 1<sup>st</sup>-trimester, 1-week prior to blood draw, and 1-month prior to blood draw) with the risk of PTB and ETB.

## Methods



### *Study population*

The current analysis included study participants enrolled in the ATL AA Maternal-Child Cohort.<sup>147,148</sup> Briefly, since 2014, this prospective cohort has recruited pregnant AA women, self-reported as U.S.-born African Americans, between 18 and 40 years of age, without chronic medical conditions and presenting for prenatal care with a singleton pregnancy estimated to be between 6 and 17 weeks of gestation at clinics of Emory Midtown Hospital (privately funded) and Grady Hospital (publicly funded). No other exclusion criteria were applied regarding pregnancy complications. Data were collected via questionnaires blood samples obtained via venipuncture at the enrollment visit (targeting 6-17 weeks). Additional details regarding recruitment and enrollment are provided elsewhere.<sup>147</sup> In total, we analyzed data from 329 women with valid metabolomics data available at the enrollment visit, enrolled between March 2014 and May 2018. This study was approved by the Emory University Internal Review Board and written informed consent was obtained from all study participants.

### *Air pollution exposure assessment*

Details of the air pollution exposure assignment and the specific model used have been previously published.<sup>149</sup> Briefly, we used ambient PM<sub>2.5</sub> concentrations at each participant's geocoded residential address provided at the first visit as the surrogate for individual exposure. The residential address was spatially joined with an ensemble-based model, which integrated multiple machine learning algorithms and predictor variables with a spatial resolution of 1km, to obtain daily PM<sub>2.5</sub> concentrations. As the exposure must precede the mediator (i.e., metabolic features), we selected four averaging periods including 1-year prior to conception, first trimester, and 1-month and 1-week prior to blood draw. The time windows except for 1-year were considered as short-term exposure.

### *Gestational age at birth outcomes*

Gestational age at birth in completed gestational weeks was abstracted from medical records and was based on the best obstetrical estimate, following the American College of Obstetrics and Gynecology (ACOG) guidelines,<sup>145</sup> considering the date of delivery in relation to the estimated date of confinement established by the 8-14-week prenatal visit..<sup>150</sup> Considering completed gestational weeks, births were classified as : PTB (>20 and <37 weeks), ETB ( $\geq 37$  and < 39 weeks), and full-term birth (FTB,  $\geq 39$  weeks).<sup>134</sup> PTB and ETB were the early birth outcomes of interest, with FTB serving as the referent category. In other words, two binary outcome variables were generated.

### *High-resolution metabolomics*

Non-fasting serum samples were analyzed using high-resolution liquid chromatography coupled with mass spectrometry (HR-LCMS, Thermo Scientific™ Q- Exactive™ HF) via established protocol<sup>147,151</sup>. Briefly, each sample [study samples and quality control (QC) samples] was run in triplicates and analyzed through two analytical columns, hydrophilic interaction liquid chromatography (HILIC) column with positive electrospray ionization (ESI) and C18 hydrophobic reversed-phase chromatography column with negative ESI. The metabolic features with mass-to-charge ratio (mz), retention time (rt), and relative intensity were extracted by R packages *apLCMS* with *xMSanalyzer*,<sup>152,153</sup> averaged, and then transformed with the natural log for downstream analysis. In addition, the relative standard deviation (RSD) was calculated for each feature across QC samples, and the missingness was estimated for study and QC samples, respectively. To tease out low-quality features but maximize metabolome coverage, we excluded features if either of these conditions was met: 1) RSD > 50% and QC missingness < 10%; 2) Missingness of study samples < 90%. As a result, 11,269 and 9,565 metabolic features remained in the current analysis for the HILIC and C18 columns, respectively. The missing values were imputed by quantile regression imputation of left-censored data (QRILC) or random forest (RF).<sup>154</sup> We classified the missing pattern [i.e., missing not at random (MNAR) vs. missing

at random (MAR)] using a second, correlated (Pearson's correlation > 0.5) auxiliary feature.<sup>155</sup>

Due to its correlation, we concluded that insights into the pattern of missing values of a given feature can be gained from the corresponding non-missing observations of its auxiliary feature.

The missing values of MNAR features were imputed by QRILC, while those of MAR features by RF, which was recommended in a previous study systematically comparing the imputation performance of different algorithms.<sup>154</sup>

### *Covariate assessment*

We determined the confounding structure based on literature review and our previous studies, which was illustrated via directed acyclic graphs (DAGs) (Figure S1). Individual-level demographic characteristics [maternal age, and maternal educational attainment (categorized as less than high school, high school, and some college or more)] were obtained via a standardized interview questionnaire. Infant sex (binary), parity (categorized as nulliparity, primiparity, and multiparity), tobacco and marijuana use in the month prior to pregnancy (binary), and alcohol use in the month prior to pregnancy (binary) were abstracted from the medical record. Maternal body mass index (BMI, kg/m<sup>2</sup>) was calculated using weight and height measured at the first visit. The meteorological covariates included the conception season (for long-term exposure) and averaged apparent temperature (for short-term exposure with the same time windows as air pollution estimates). The daily apparent temperature at the metro Atlanta airport were obtained from Automated Surface Observing System via R package *riem*.<sup>156</sup>

### *Statistical analysis*

We summarized maternal and newborn characteristics for women stratified by the gestational age at birth categories of interest. We tabulated the arithmetic means and standard deviations (SDs) of exposures and apparent temperature averages (Table 2, Figure S2), and a descriptive statement of exposures was included in the main text. To identify the potential metabolic

features mediating the association of  $PM_{2.5}$  with PTB and ETB, we adopted a parallel strategy using the Meet-in-the-Middle (MITM) and high-dimensional mediation analysis (HDMA) approaches simultaneously (Figure 1).

Firstly, we conducted a metabolome-wide association study (MWAS) for exposures and outcomes separately and followed a MITM approach to identify the overlapping features associated with both exposures and outcomes. MITM is a widely-used step in high-dimensional settings to identify intermediate biomarkers.<sup>93</sup> Specifically, we conducted a series of multiple linear regression (i.e., exposure-mediator) models and logistic regression (mediator-outcome) models to evaluate the association of metabolic features with exposures and outcomes, respectively, using the following equations:

$$\ln(\text{Feature}_j) = \beta_{0j} + \beta_{1j}PM_{2.5} + \beta_{2j}Age + \beta_{3j}Education + \beta_{4j}Sex + \beta_{5j}BMI + \beta_{6j}MET + \beta_{7j}Toba\_mari + \beta_{8j}Alcohol + \epsilon_j \quad \text{Eq. (1)}$$

$$\text{Logit}(P(\text{Birth})) = \theta_{0j} + \theta_{1j} \ln(\text{Feature}_j) + \theta_{2j}PM_{2.5} + \theta_{3j}Age + \theta_{4j}Education + \theta_{5j}Sex + \theta_{6j}BMI + \theta_{7j}Parity + \theta_{8j}Toba\_mari + \theta_{9j}Alcohol \quad \text{Eq. (2)}$$

where  $\ln(\text{Feature})$  refers to the natural log of intensity of metabolic feature  $j$ ;  $PM_{2.5}$  is the averaged  $PM_{2.5}$  exposure for a specific window, and  $MET$  is the corresponding meteorological covariate;  $Birth$  denotes PTB and ETB, and the FTB group was treated as reference; we included  $PM_{2.5}$  in the mediator-outcome model to block the direct effect of  $PM_{2.5}$  exposure on the outcome, which may confound the mediator-outcome association. We constructed the two equations following a sophisticated causal mediation framework with adjustment of exposure-mediator confounders and mediator-outcome confounders (Figure S1) in the Eq. (1) and Eq. (2), respectively.<sup>157</sup> We used Benjamini-Hochberg adjusted p-values ( $FDR_{B-H}$ )  $< 0.2$  as the cut-off point for significance.<sup>158</sup> To reduce Type II error and maximize the detection of potential mediators, we also explored the overlapping features among the top 100 features ranked by raw  $p$ -values. Results were presented using Manhattan plots (Figure S3 & S4 in the Supplement).

As a complementary means to identify the potential mediating features between PM<sub>2.5</sub> exposures and the gestational age at birth outcome categories, we employed HDMA via the R package *HIMA*.<sup>94</sup> Previous researchers have developed a framework of mediation analysis that is able to deal with multiple mediators simultaneously and tease apart the indirect effect of individual mediator, which was depicted elsewhere in details.<sup>159</sup> *HIMA* expands this multiple mediator framework to the high-dimensional setting by reducing the dimensionality of omics data, and the significant mediators were reported with multiple testing correction.<sup>94</sup> Compared to the aforementioned MITM approach, *HIMA* is able to incorporate multiple mediators in a single mediator-outcome model, which enables us to ascertain the extent to which the indirect effects are explained by the mediators. Separate analysis was conducted for each column (HILIC positive ESI and C18 negative ESI).

To aid the interpretation of the MITM approach and HDMA results, we conducted a pathway enrichment analysis using the R package *metapone* using raw *p*-values of metabolic features at 0.05 in both MWAS. *Metapone* is a novel bioinformatic platform to predict functional biological activities of untargeted metabolomic data extracted in both positive and negative ESI together, which developed a pathway database combining the Small Molecule Pathway Database (SMPDB) and mummichog database. The inputs of metabolic features were putatively annotated with the related weights calculated based on the uncertainty in metabolite-feature matching, and then the significance of enriched biological pathways was tested taking into account the weight schema.<sup>160</sup> The biological pathways associated with either PM<sub>2.5</sub> exposures or outcomes with more than one metabolite enriched and a *p*-value < 0.05 were included for further detecting the overlapping pathway.

The metabolic features associated with both PM<sub>2.5</sub> exposure and gestational age at birth outcomes were identified and annotated for functional interpretation. We confirmed the identity of overlapping features by comparison of *m/z*, *rt*, and ion dissociation patterns to authentic chemical reference (confidence level 1) or the annotation procedure of the R package

*xMSannotator* (confidence level 4).<sup>161</sup> The confidence system referred to the a five level system of reporting standard proposed by Schymanski et al. with, for example, confidence 1 representing the proposed chemical identity that is confirmed via comparing to an authentic standard.<sup>162</sup> All analyses were completed in R (version 3.6).

## Results

A total of 329 individuals from the Atlanta AA cohort were included in the current analysis and their demographic characteristics are described in Table 1, stratified by gestational age at birth outcome category. Participants with a PTB or ETB reported less education compared to those with FTB. Participants with a PTB had the lowest early pregnancy BMI, highest proportion of multiparity, highest infant sex ratio (Male vs. Female  $\approx$  3:2), and highest proportion of maternal tobacco and marijuana use; while those with ETB had the lowest proportion of multiparity, highest proportion of maternal alcohol use, and other covariates with a similar distribution compared to those with an FTB.

The median of PM<sub>2.5</sub> exposure during the one-year prior to conception, first trimester, one-week and one-month prior to blood draw were 10.39 [interquartile range (IQR)=1.15], 10.31 (2.63), 9.95 (3.87), and 10.17 (3.11)  $\mu\text{g}/\text{m}^3$ , respectively (Table 2). The long-term exposure (i.e., one-year average) was not correlated with the three short-term exposures, whereas the short-term exposures were moderately to strongly correlated with each other (Figure S2A).

Participants living in urban areas had a higher level of one-year exposure compared to those in suburban areas, but we did not observe the same tendency for the short-term exposure (Figure S2B).

### *Metabolome-wide association analysis*

There were 17 metabolic features associated with 1-year exposure prior to conception, 3 features with 1<sup>st</sup> trimester, 17 features with 1-week prior to blood draw, and 1 feature with 1-month prior to blood draw in the HILIC column ( $FDR_{B-H} < 0.2$ ). Similarly, 2 metabolic features were found to be associated ( $FDR_{B-H} < 0.2$ ) with 1<sup>st</sup> trimester, 20 features with 1-week, and 1 feature with 1-month exposure in the C18 chromatography column. The detailed statistics of significant metabolic features were summarized in the Supplement (Table S1 & S2). Two features ( $mz$ : 268.1082,  $rt$ : 81.9;  $mz$ : 282.1185,  $rt$ : 119.1) associated with one-week exposure were confirmed as adenosine and methyladenosine, respectively, with level 1 confidence. Methyladenosine is a modified form of adenosine where a methyl group is added and is the most common modifications found in mRNA.<sup>163</sup> One feature ( $mz$ : 391.2842,  $rt$ : 22) associated with both 1<sup>st</sup>-trimester and one-month exposures was confirmed as Di(2-ethylhexyl)phthalate (DEHP) which is a frequently used plasticizer and has a profound impact on human health.<sup>164</sup> We did not find any significant features in the mediator-outcome (i.e., metabolite-PTB/ETB) models after multiple testing correction.

### *Meet-in-the-Middle approach*

In order to identify metabolic features that potentially mediated the association of  $PM_{2.5}$  exposures with PTB and ETB, we employed different significance cutoff to detect the overlapping features associated with  $PM_{2.5}$  exposures and PTB and ETB, simultaneously (Table S3). We subset the top 100 features sorted by raw  $p$ -values that were associated with any exposure and outcome, and examined the overlapping features between exposures and outcomes in both columns (Figure 2, Table S3). Eleven features were identified as overlapping, and seven had the same direction for both coefficients of exposure-mediator, in this case  $PM_{2.5}$ , and mediator-outcome models, indicative of the mediation of adverse effect of  $PM_{2.5}$  exposures on PTB and ETB. At raw  $p$ -value  $< 0.05$ , the number of overlapping features had a range from 13 to 99 (Median=25.5) among the associations between four exposure windows and two

adverse outcomes (Table S3). Of note, there were 99 overlapping features potentially mediating the association between one-year exposure to PM<sub>2.5</sub> prior to conception and ETB. Among the annotated overlapping features (Table S4), N6,N6,N6-Trimethyl-L-lysine, a metabolite involved in lysine degradation, was associated with 1<sup>st</sup> trimester exposure and PTB; N-Acetyl-D-Galactosamine (Galactose metabolism) was associated with 1<sup>st</sup> trimester, one-week, one-month exposures and PTB; Cortisolone (Cortisol synthesis and secretion) was associated with one-week exposure, PTB, and ETB; Picolinic acid and serotonin, both belonging to tryptophan metabolism, were associated with one-year exposure and ETB; Citrulline (Arginine biosynthesis), which was detected by both HILIC and C18 columns, was associated with one-year exposure and ETB.

#### *High-dimensional mediation analysis*

Five features were identified using HDMA, among which three had a positive indirect effect estimate (Figure 2). Three features were detected by both MITM (Top 100) and HDMA ( $FDR_{B-H} < 0.05$ ) in the HILIC and C18 columns [arbitrarily labelled as feat\_1 (mz: 202.0862; rt: 227.2), feat\_2 (mz: 1086.4312, rt: 276.9), and feat\_3 (mz: 445.1996; rt: 216.4)]. One feature (mz: 570.5164, rt: 47.7) was also detected by the MITM approach at the significance level of raw  $p$ -values  $< 0.05$  to be associated with one-year exposure and PTB. We were not able to confirm the identity of any of these features with authentic reference standards (level 1 confidence). In sum, eight metabolic features detected by either MITM (Top 100) or HDMA ( $FDR_{B-H} < 0.05$ ) potentially mediated the adverse effect of prenatal PM<sub>2.5</sub> exposures on PTB and ETB.

#### *Pathway enrichment analysis*

Five biological pathways were found to potentially mediate the association with ETB (Figure 3): four for 1-year exposure prior to conception, one for the 1-week exposure, and one overlapping pathway was identified for PTB associated with the 1-week exposure. The pathways included



membrane transport [ATP-binding cassette (ABC) transporters], amino acid metabolism (glycine, serine, alanine, and threonine metabolism), and vitamin B9 (folate) metabolism which is vital for fetus growth and pregnancy.

## Discussion

In the well-established and highly phenotyped Atlanta African American Maternal Child cohort, we identified eight metabolic features, which were detected in maternal serum metabolome during early pregnancy, with potential important roles in mediating the adverse effects of PM<sub>2.5</sub> exposures on the risk of PTB and ETB. We also observed perturbations in several amino acid and one carbon metabolic pathways that were associated with both PM<sub>2.5</sub> and ETB.

We employed a parallel strategy by applying both the MITM approach and HDMA to evaluate the mediating role of metabolomic perturbations on the association between PM<sub>2.5</sub> and adverse birth outcomes. In the current analysis, the MITM approach based on MWAS is capable of detecting potential intermediate factors (i.e., the overlapping metabolic features) at various significance levels lying on a potentially biological pathway from PM<sub>2.5</sub> exposure to PTB or ETB and enables the posterior identification of overlapping biological pathways which facilitates biological interpretation of untargeted metabolomic data. Furthermore, HDMA incorporated multiple metabolic features into a single mediator-outcome model, which teased apart the indirect effect of PM<sub>2.5</sub> exposure on PTB and ETB and estimated the indirect effect associated a particular feature independent of other features. HDMA might be more in line with reality in situations where the relationship between PM<sub>2.5</sub> exposure and PTB and ETB is complex and involves multiple biological mechanisms.

Our study findings provide novel insights on how biological perturbations, centering around amino acid metabolism during early pregnancy may mediate the adverse effects of PM<sub>2.5</sub> exposure on preterm or early-term birth among pregnant AA people. The mechanistic insight

may contribute to the future development of sensitive biomarkers predicting PM<sub>2.5</sub>-related adverse birth outcome for effective intervention. Greater intervention may, in turn, ultimately alleviate health disparities associated with PM<sub>2.5</sub> exposures among pregnant women and newborns. The eight metabolic mediators found to mediate PM<sub>2.5</sub> toxicity, contributing evidence of a potential link between prenatal PM<sub>2.5</sub> exposure and PTB and ETB.

We identified seven metabolic mediators using the MITM approach from the top 100 metabolic features of the MWAS for PM<sub>2.5</sub> exposures, PTB and ETB. To maximize the potential to detect the mediators, we further used the significance level of raw *p*-values < 0.05 and identified a range of 13-99 overlapping features for different exposure-outcome associations, and a few features were annotated with level 1 confidence as N6,N6,N6-Trimethyl-L-lysine, N-Acetyl-D-Galactosamine, cortexolone, picolinic acid, serotonin, and citrulline. Cortexolone, also known as 11-deoxycortisol, is an endogenous glucocorticoid steroid and a metabolic intermediate in the synthesis of cortisol which is a potential contributor to premature labor.<sup>165</sup> In previous literature, an increased serum level of cortexolone was observed among male Sprague-Dawley rats exposed to carbon black nanoparticles for 90 days,<sup>166</sup> while in a population-based cohort study of 6,670 Chinese rural residents, no significant association was observed between long-term residential exposure to PM<sub>2.5</sub> and serum cortexolone level.<sup>167</sup> On the other hand, the serum cortexolone level in the 3<sup>rd</sup> trimester was found to be positively associated with a higher risk of spontaneous premature birth (< 32 gestational weeks) in a prospective pregnancy biorepository in Indiana state.<sup>168</sup> Citrulline, a non-essential amino acid, serves a fundamental role in the urea cycle along with arginine. Hu et al. (2021), conducted an exposure-interventional study of 43 healthy adults in which an air purifier was installed in participants' office rooms and dormitories, and a positive association was observed between the short-term personal exposure to PM<sub>2.5</sub> and plasma citrulline level, suggesting acute perturbations of amino acid metabolism.<sup>169</sup> Citrulline supplementation was demonstrated to enhance fetal growth, protein synthesis, and placental function among rats with intrauterine

growth restriction, while human studies are scarce.<sup>170</sup> More research is warranted on steroid hormone metabolism and amino acid metabolism to investigate their potential role in the PM<sub>2.5</sub>-related PTB and ETB. In the current study, three metabolic features were detected via HDMA to mediate a positive association of prenatal PM<sub>2.5</sub> exposures with PTB and ETB, and one metabolic feature (mz: 202.0862, rt: 227.2) mediated the association between 1-month exposure prior to blood draw, which was consistently detected by MITM and HDMA. Although none of the three features were annotated with level 1 confidence, the consistencies between MITM and HDMA suggest novel biomarkers that might lead to advances in the development of new interventions and prediction of birth outcomes.

Our analysis adds to a growing body of research investigating underlying mechanisms of air-pollution-related adverse birth outcome by characterizing maternal metabolomic profiles.<sup>27</sup> Previous studies that examined the associations of the maternal metabolomics perturbations with air pollution exposure found that the levels of serum serine, creatinine, histidine, myo-inositol, heptadecanoic acid, and linoleic acid in mid pregnancy were associated with high traffic-related air pollution during the first trimester among women of multiple races;<sup>141</sup> five steroids in mid- and late-pregnancy were associated with PM<sub>2.5</sub> exposure during early pregnancy in an ethnically diverse population.<sup>171</sup> Moreover, maternal metabolomic profiles might be capable of predicting birth outcomes, such as gestational age, time to delivery,<sup>172</sup> and term fetal growth restriction.<sup>173</sup>

Another key finding from our analysis was the potential role that perturbed amino acid metabolism played in mediating the adverse effects of prenatal PM<sub>2.5</sub> on early birth. In particular, we identified six overlapping biological pathways associated with PM<sub>2.5</sub> exposures and PTB and ETB, some of which were also reported in previous literature to be associated with air pollution exposure.<sup>141,142,146,174-178</sup> Specifically, glycine, serine, alanine and threonine metabolism were found to be associated with short-term exposure to black carbon, PM<sub>2.5</sub>, and distance to major roadways among women undergoing assisted reproduction.<sup>146,174</sup> Yan et al. (2019) reported the

glycine, serine, alanine and threonine metabolism associated with elevated traffic-related air pollution exposure in maternal serum metabolome during early pregnancy.<sup>141</sup> Vitamin B9 (folate) metabolism was found by pathway enrichment analysis in multiple previous studies among various study populations exposed to particulate matter and its species.<sup>142,177-180</sup> Vitamin B9, also known as folate, is an essential nutrient that plays a critical role in many physiological processes, and one of the most important roles of folate is in the synthesis of DNA and RNA.<sup>181</sup> Folate is also involved in the metabolism of amino acids, such as homocysteine, and important for proper fetal development.<sup>182</sup>

Collectively, these findings point to the vital roles of amino acid metabolism, centering around the urea cycle, creatine pathway, proline biosynthesis, histidine degradation, serine and cysteine biosynthesis, in mediating the effects of prenatal PM<sub>2.5</sub> exposure on adverse birth outcomes (Figure 4). Of note, citrulline, as an essential part of urea cycle, was detected by the MITM approach in the current analysis. Amino acid metabolism broadly refers to the process by which the body synthesizes, metabolizes, and transforms various amino acids. Homeostatic balance of these molecules is especially essential during pregnancy, a unique life stage defined by significant maternal physiologic changes and rapid fetal growth.<sup>183</sup> The metabolomic changes we observed as well as those reported previously, point specifically to perturbations in certain submodules of amino acid metabolism, and may elucidate the specific mechanistic pathway underlying the association between PM<sub>2.5</sub> exposure and these adverse birth outcomes. As we hypothesize in Figure 4, serine, a non-essential amino acid, plays a critical role in carbon cycling of human body. It is a major source of one specific carbon pool, required for glycine-based formation of cysteine, glucose, and other key regulating molecules in utero<sup>184</sup>. A lower level of plasma serine and rate of appearance were reported among healthy pregnant women compared to nonpregnant women,<sup>185</sup> possibly implying high turnover of serine in the fetal compartment.<sup>184</sup>

Formate was enriched for PM<sub>2.5</sub> exposure and PTB and ETB in the current analysis, which is a key intermediate in one carbon metabolism. Cumulative evidence suggested a critical role in pregnancy and fetus development. A high level of urine formate was significantly associated with a decreased incidence of fetal growth restriction,<sup>186</sup> and the formate concentrations in maternal plasma were found to be lower than that in cord blood among 215 mothers at delivery, indicating a higher activity of one-carbon metabolism in fetuses.<sup>187</sup> Although evidence is scarce on the effect of PM<sub>2.5</sub> exposure on formate metabolism, in a study conducted by Gaskins and colleagues in 2019, a potential modifying effect of folate intake on the association between short-term exposure to air pollution and the odds of livebirth was observed among women undergoing assisted reproduction.<sup>188</sup> Taken these together, we hypothesized that elevated exposures to PM<sub>2.5</sub> during early pregnancy and prior to conception would result in perturbations in urea cycle and serine and cysteine metabolism, leading to the dysregulation of already fragile amino acid balance during pregnancy, which potentially contributed to enhanced protein-energy undernutrition, eventually leading to increased risk of PTB and ETB.

Our study has several notable strengths in addition to the novel strategy using both the MITM approach and HMDA. Gestational age was well characterized in early pregnancy.<sup>147</sup> The analytical cohort we analyzed was exclusively AA, a population that has been largely underrepresented in environmental epidemiologic studies. Finally, the workflow of the untargeted metabolomics profiling was well established and has been shown to successfully analyze many non-fasting samples previously.<sup>146,189</sup>

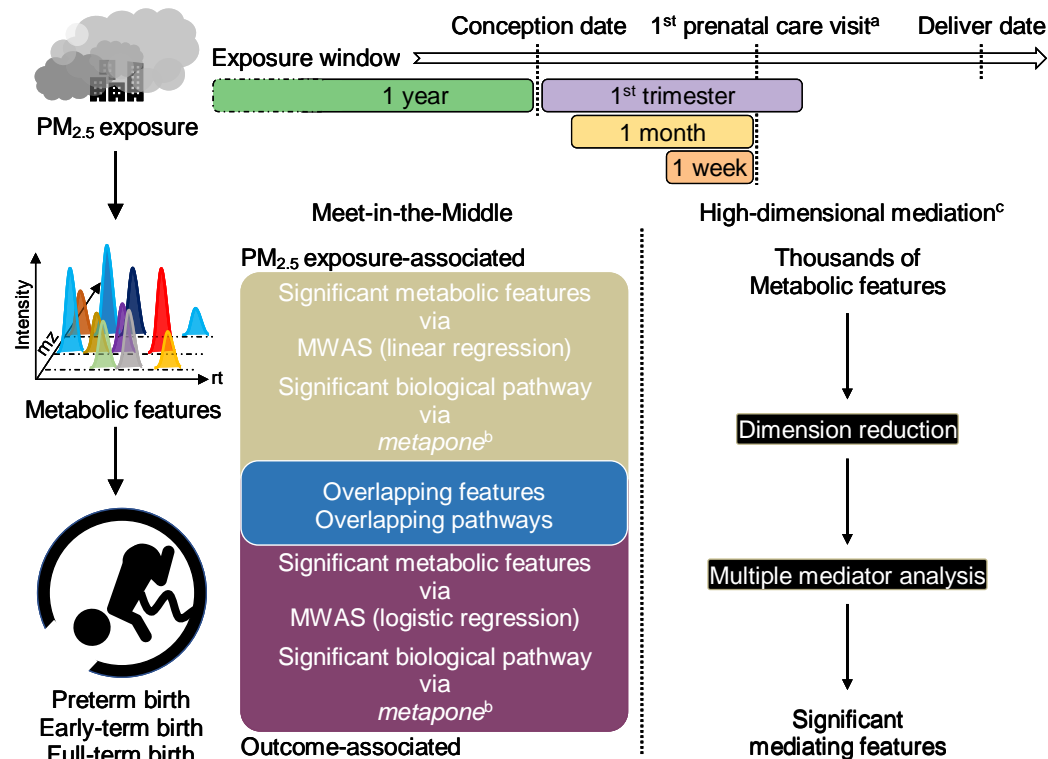
Despite these strengths, we identified several key limitations and areas for future work. First, our results may not necessarily imply a causal relationship. These alterations could potentially be due to the presence co-exposure confounding, which should be explored in future mixtures work. Second, the serum samples were collected in early and mid-pregnancy and we were, thus, not able examine exposures occurring in other potentially critical exposure time windows.

Third, ambient PM<sub>2.5</sub> exposure was estimated based on the residential address, which did not consider daily mobility patterns.

## Conclusions

Using innovative high-resolution metabolomics and mediation analysis, we identified novel metabolomics signatures, centering around the amino acid metabolism and one carbon metabolism, mediating the adverse effects of PM<sub>2.5</sub> exposure on the risk of PTB and ETB in a marginalized and understudied population. Our findings provide important information about the metabolic perturbations that appear to mediate the association between PM<sub>2.5</sub> and PTB and ETB. This information is useful for understanding the biological mechanisms underlying PM<sub>2.5</sub> toxicity on pregnancy outcomes and lend support to future development of biomarker profiles that might facilitate the identification and potential targeting of interventions for those at elevated risk.

## Tables and figures



**Figure 1.** Graphical overview of the parallel analysis strategy to evaluate the potential mediating metabolic features in the current analysis. <sup>a</sup> The date of first prenatal care visit varied by participating mothers from 8 to 14 gestational weeks. <sup>b</sup> *Metapone* was an R package to conduct pathway enrichment analysis for untargeted metabolomics data. <sup>c</sup> Conducted by the R package *HIMA*. Mz, mass-to-charge ratio; rt, retention time; MWAS, metabolome-wide association study.

**Table 1.** Selected population characteristics by birth outcomes among the subjects enrolled in Atlanta African American Maternal-Child Cohort study, 2014-2018 (N = 329).

	Preterm (N=65)	Early-term (N=54)	Full-term (N=210)	<i>p</i> -value
Maternal age, mean (SD)	24.4 (4.56)	25.1 (4.84)	25.1 (4.84)	0.592
Maternal educational attainment, No. (%)				
Less than high school	12 (18.5)	14 (25.9)	28 (13.3)	0.042
High school	30 (46.2)	20 (37.0)	73 (34.8)	
Some college or more	23 (35.4)	20 (37.0)	109 (51.9)	
Body mass index, mean (SD)	27.2 (6.91)	28.1 (8.17)	29.2 (7.68)	0.185
Parity, No. (%)				
Nulliparity	28 (43.1)	16 (29.6)	104 (49.5)	0.022
Primiparity	13 (20.0)	22 (40.7)	55 (26.2)	
Multiparity	24 (36.9)	16 (29.6)	51 (24.3)	
Infant sex, No. (%)				
Male	40 (61.5)	26 (48.1)	98 (46.7)	0.125
Female	25 (38.5)	28 (51.9)	112 (53.3)	
Maternal tobacco or marijuana use, No. (%)				
No	36 (55.4)	33 (61.1)	119 (56.7)	0.810
Yes	29 (44.6)	21 (38.9)	91 (43.3)	
Maternal alcohol use, No. (%)				
No	59 (90.8)	48 (88.9)	189 (90.0)	0.937
Yes	6 (9.2)	6 (11.1)	21 (10.0)	
Season of conception, No. (%)				
Spring	19 (29.2)	12 (22.2)	52 (24.8)	0.930
Summer	21 (32.3)	18 (33.3)	67 (31.9)	
Fall	15 (23.1)	12 (22.2)	45 (21.4)	
Winter	10 (15.4)	12 (22.2)	46 (21.9)	

Abbreviations: SD, standard deviation.

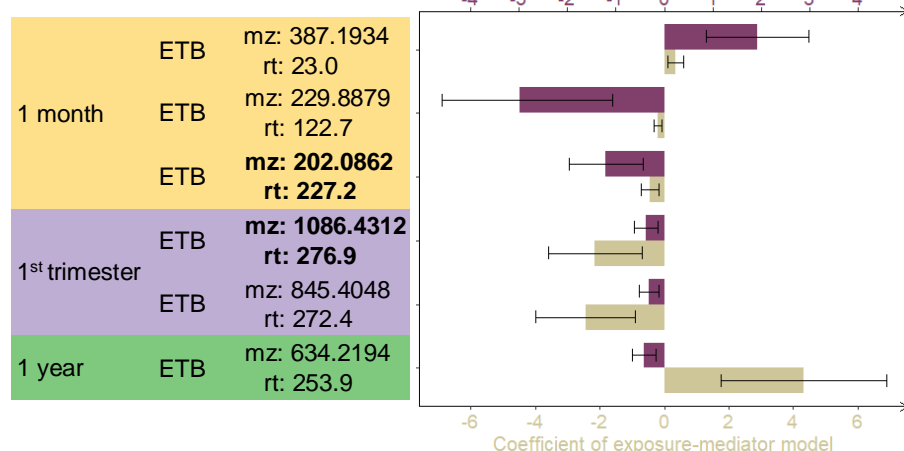


**Table 2.** Statistics of fine particulate matter (PM<sub>2.5</sub>) exposure for the four exposure windows among 329 pregnant people in the Atlanta African American Maternal-Child Cohort, 2014-2018.

PM <sub>2.5</sub> exposure (µg/m <sup>3</sup> )	Q1	Q3	IQR	Median	Mean
One-year exposure prior to conception	9.80	10.95	1.15	10.39	10.34
Exposure for the 1 <sup>st</sup> trimester	9.13	11.75	2.63	10.31	10.58
One-week exposure prior to blood draw	8.19	12.07	3.87	9.95	10.38
One-month exposure prior to blood draw	8.64	11.74	3.11	10.17	10.49

Abbreviations: Q1, 1<sup>st</sup> quartile; Q3, 3<sup>rd</sup> quartile; IQR, interquartile range.

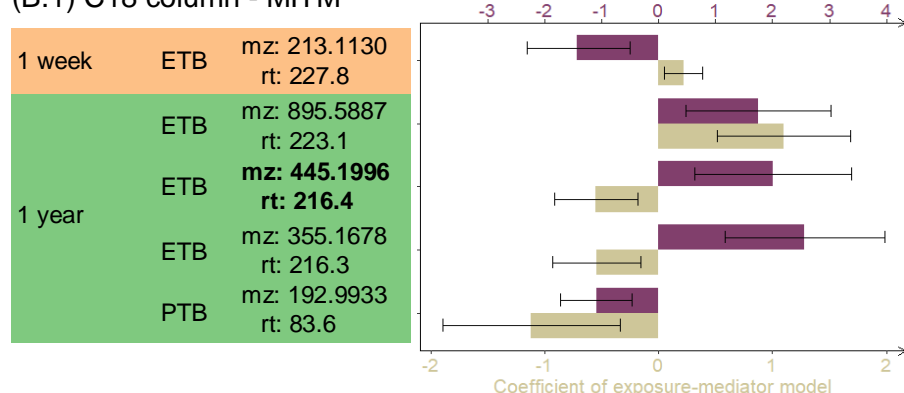
## (A.1) HILIC column - MITM



## (A.2) C18 column - HDMA

			Indirect effect estimates	FDR <sub>B-H</sub>
1 month	ETB	mz: 1086.431 rt: 276.9	0.31	0.003
		mz: 202.0862 rt: 227.2	0.38	0.003
1 year	PTB	mz: 570.5164 rt: 47.7	1.06	0.023
		mz: 444.3315 rt: 22.7	-0.72	0.035

## (B.1) C18 column - MITM

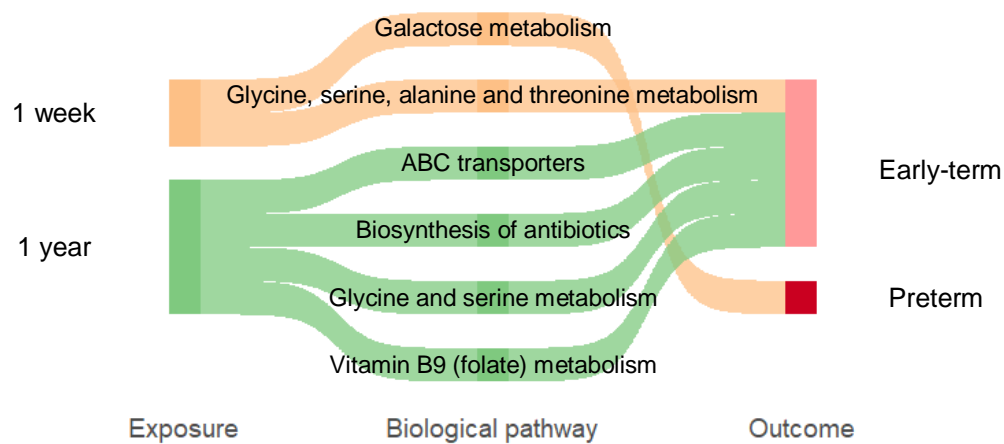


## (B.2) C18 column - HDMA

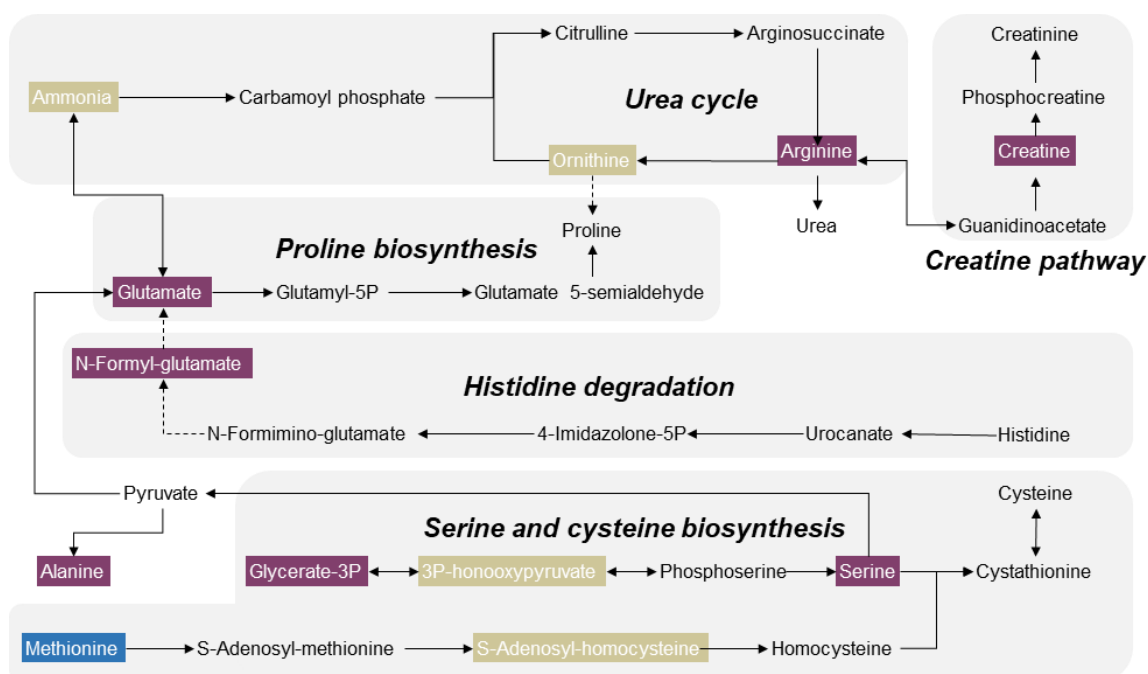
			Indirect effect estimates	FDR <sub>B-H</sub>
1 year	ETB	mz: 445.1996 rt: 216.4	-0.82	0.004

**Figure 2.** The overlapping and mediating metabolic features detected by the Meet-in-the-Middle (MITM) approach and high-dimensional mediation analysis (HDMA), respectively. Panel (A) shows the results of the HILIC column, and panel (B) shows the C18 column. In each panel, the upper plot (A.1 and B.1) shows the MITM results, while the lower shows (A.2 and B.2) the HDMA results. The upper plot shows the coefficients and unadjusted 95% confidence interval of each overlapping features estimated in the exposure-mediator and mediator-outcome models. The coefficient for the exposure-mediator model represented the change in log-transformed feature intensity associated with 10  $\mu\text{g}/\text{m}^3$  increase in  $\text{PM}_{2.5}$  levels. The lower plot shows indirect effect estimate of each mediating feature. The indirect effect was also associated with 10- $\mu\text{g}/\text{m}^3$  increase in  $\text{PM}_{2.5}$  exposure. Metabolic features were annotated with level 4 confidence, and those not annotated were listed with mz and retention time. ETB, early-term birth; PTB, preterm

birth; mz, mass-to-charge ratio; rt, retention time;  $\text{FDR}_{\text{B-H}}$ ,  $p$ -values adjusted by Benjamini-Hochberg (BH) procedure.

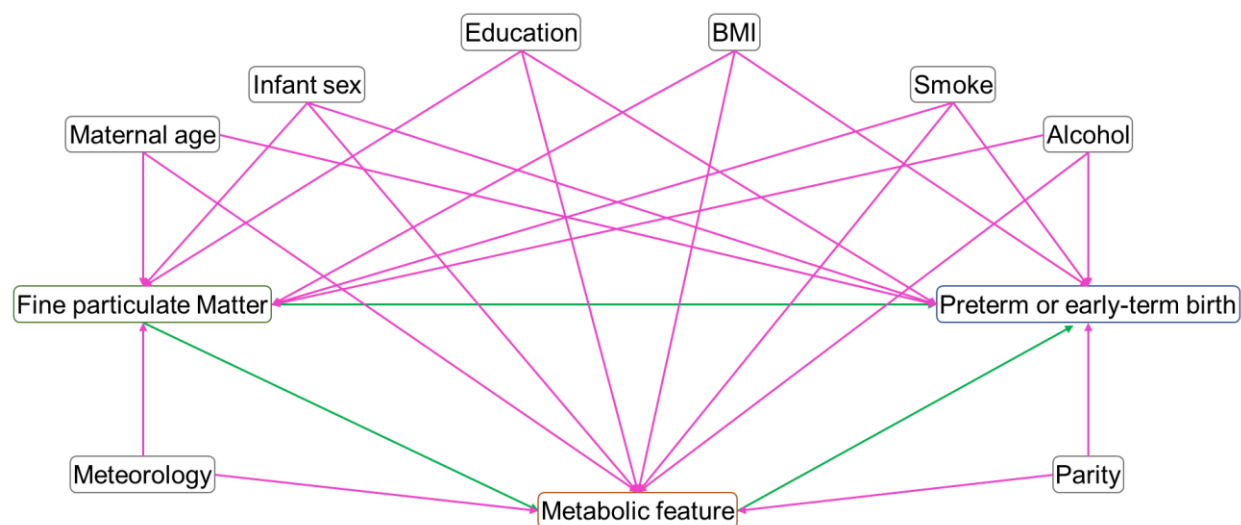


**Figure 3.** The Sankey diagram of overlapping biological pathways detected by the Meet-in-the-Middle approach coupled with *metapone*. No overlapping pathways were identified between PM<sub>2.5</sub> exposures and preterm birth.

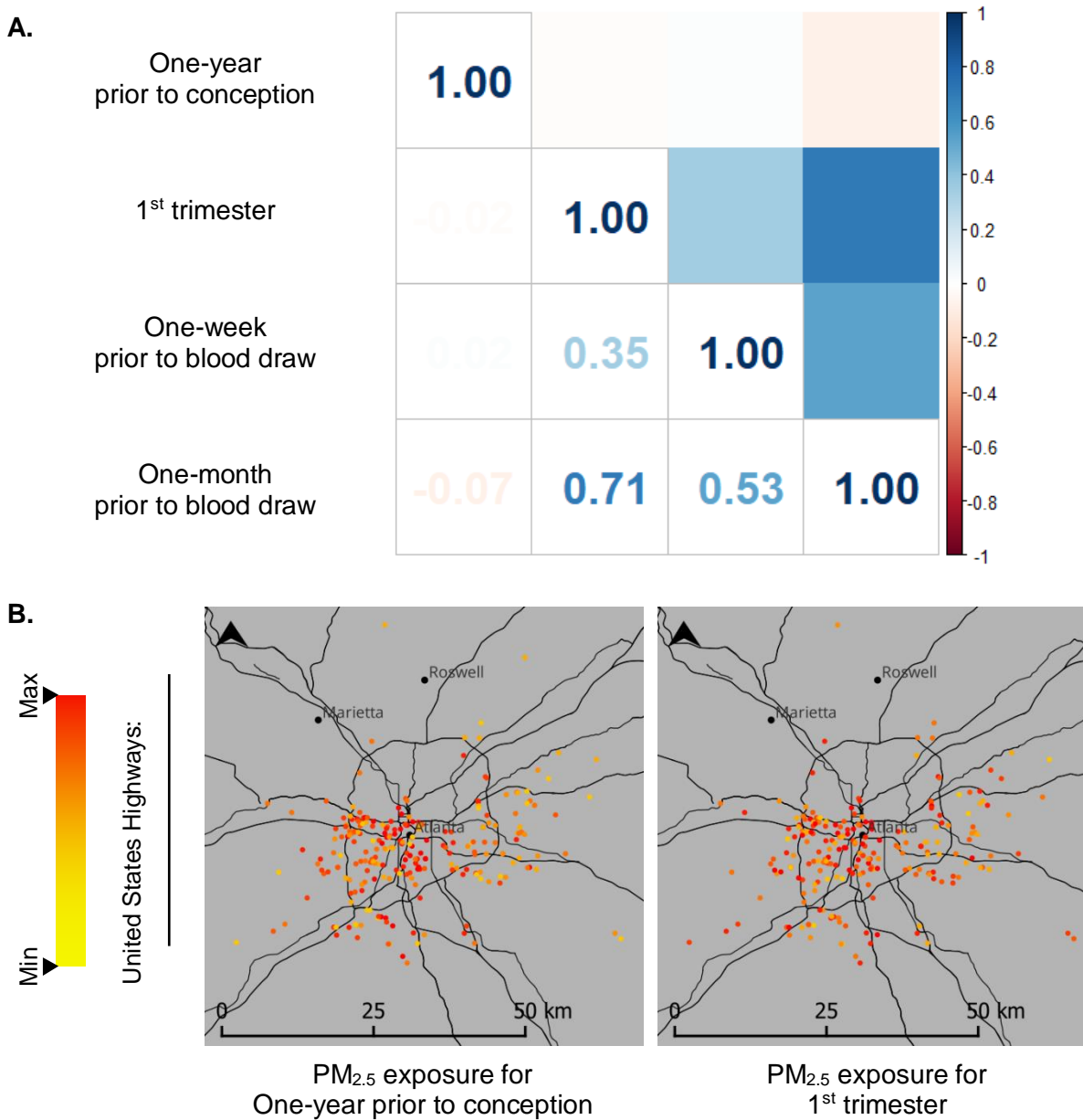


**Figure 4.** Potential molecular mechanisms as illustrated as metabolic networks for the adverse effects of PM<sub>2.5</sub> exposures on early-term birth among pregnant African American women. Molecules in martini olive, dark violet, and bice blue denoted the metabolites putatively annotated by *metapone* that were significantly associated with PM<sub>2.5</sub> exposures, ETB, and both, respectively. Dash arrows denote the reactions found in non-human species.

## Appendix for Chapter 4

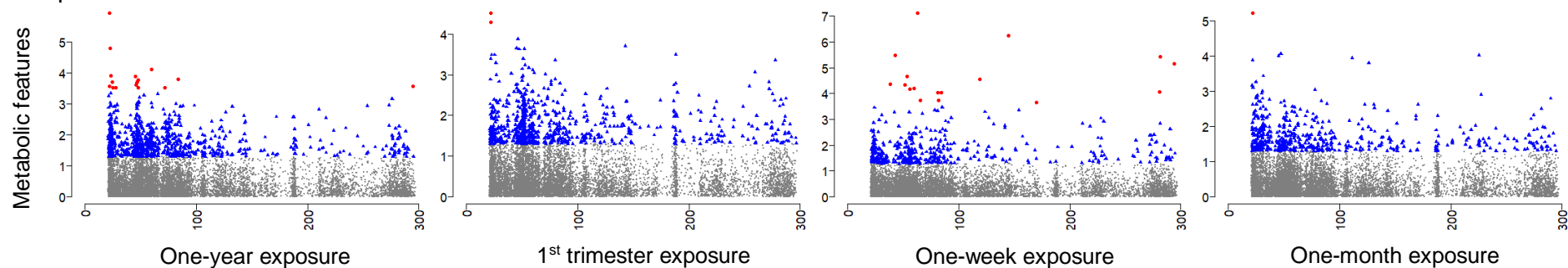


**Figure S1.** Directed acyclic graph of the confounding structure for the association of fine particulate matter exposure with preterm and early-term birth. BMI, body mass index; Smoke, tobacco and marijuana use in the month prior to pregnancy; Meteorology, conception season or averaged apparent temperature.

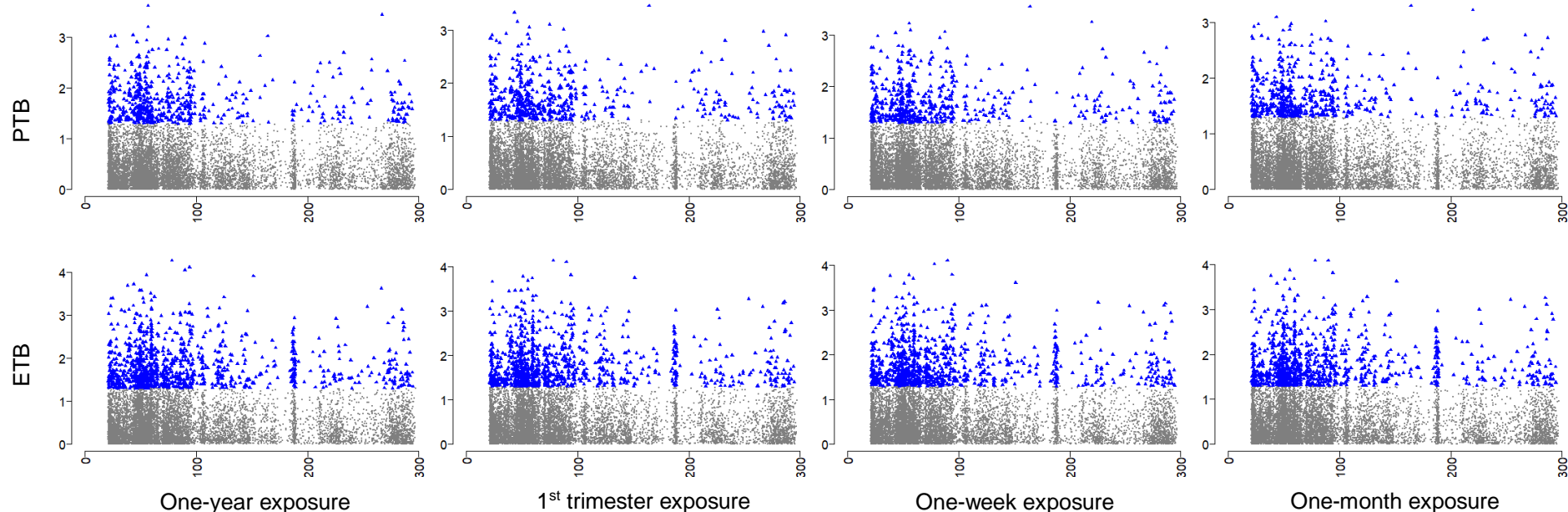


**Figure S2.** The correlations among fine particulate matter (PM<sub>2.5</sub>) exposures for the four exposure windows and the spatial variability of the PM<sub>2.5</sub> illustrated in Metropolitan Atlanta.

### A. Exposure-mediator metabolome-wide associations



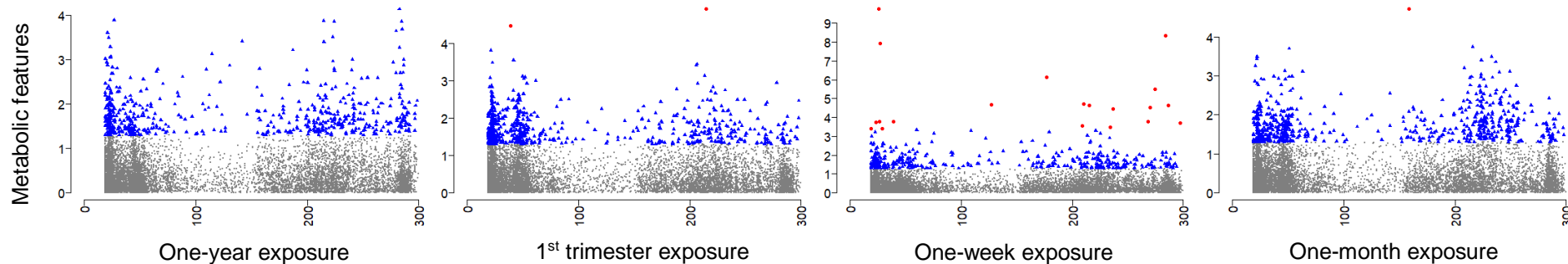
### B. Mediator-outcome metabolome-wide associations



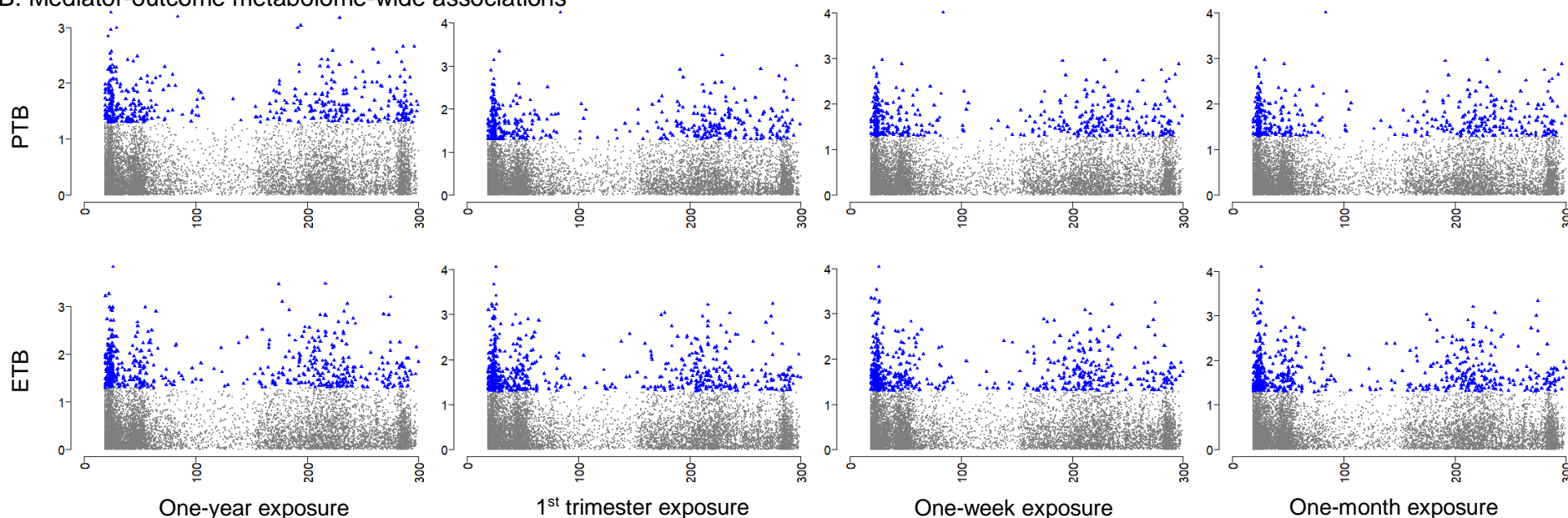
**Figure S3.** Manhattan plots of metabolome-wide association analysis in the HILIC column. A. Associations between  $PM_{2.5}$  exposures and changes in intensities of metabolic features; B. Associations between changes in intensities of metabolic features and PTB or ETB, with adjustment of  $PM_{2.5}$  exposures for different time windows separately. X-axis denotes the retention time (in seconds) of the metabolic features, and Y-axis denotes the negative  $\log_{10}$  of  $p$ -values. Red dots indicated significant associations at  $FDR_{B-H} < 0.2$ , and blue indicated associations at raw  $p$ -values  $< 0.05$ . Abbreviations: HILIC, hydrophilic interaction liquid chromatography;  $PM_{2.5}$ , fine particulate matter; PTB, preterm birth; ETB, early-term birth;  $FDR_{B-H}$ , Benjamini-Hochberg adjusted  $p$ -values.



### A. Exposure-mediator metabolome-wide associations



### B. Mediator-outcome metabolome-wide associations



**Figure S4.** Manhattan plots of metabolome-wide association analysis in the C18 column. A. Associations between PM<sub>2.5</sub> exposures and changes in intensities of metabolic features; B. Associations between changes in intensities of metabolic features and PTB or ETB, with adjustment of PM<sub>2.5</sub> exposures for different time windows separately. X-axis denotes the retention time (in seconds) of the metabolic features, and Y-axis denotes the negative log<sub>10</sub> of *p*-values. Red dots indicated significant associations at FDR<sub>B-H</sub> < 0.2, and blue indicated associations at raw *p*-values < 0.05. Abbreviations: C18, hydrophobic reversed-phase chromatography; PM<sub>2.5</sub>, fine particulate matter; PTB, preterm birth; ETB, early-term birth; FDR<sub>B-H</sub>, Benjamini-Hochberg adjusted *p*-values.

**Table S1.** Model statistics of metabolic features in the HILIC column associated ( $FDR_{B-H} < 0.2$ ) with ambient  $PM_{2.5}$  exposures of different time windows among 329 pregnant people in the Atlanta African American Maternal-Child Cohort, 2014-2018.

Exposure	mz	RT	Coefficients <sup>a</sup>	Percent change <sup>b</sup> (95% CI)	p-values	$FDR_{B-H}$
One-year	150.0267	25.1	-0.50	-39.15 (-53.33, -20.65)	2.99E-04	0.1999
	200.877	295	0.08	8.65 (3.98, 13.53)	2.64E-04	0.1999
	293.0512	45.2	0.24	26.75 (12.49, 42.81)	1.27E-04	0.1999
	338.2417	27.3	0.08	8.77 (3.99, 13.77)	3.02E-04	0.1999
	366.1863	83.6	0.34	41.04 (18.32, 68.13)	1.58E-04	0.1999
	<b>392.2877</b>	<b>22</b>	-0.65	-47.85 (-59.62, -32.65)	1.10E-06	0.0129
	429.2403	47	-0.17	-15.60 (-22.65, -7.89)	1.79E-04	0.1999
	444.3315	22.7	0.17	19.01 (10.14, 28.59)	1.57E-05	0.0887
	488.3578	22.8	0.15	15.97 (7.66, 24.91)	1.21E-04	0.1999
	500.5785	47.9	0.25	27.73 (12.11, 45.52)	2.88E-04	0.1999
	515.6465	46.1	0.21	23.79 (10.67, 38.46)	2.35E-04	0.1999
	568.5189	46.6	0.12	13.11 (6.10, 20.58)	2.00E-04	0.1999
	570.5164	47.7	0.14	14.53 (6.85, 22.76)	1.61E-04	0.1999
	646.9514	24.2	-0.07	-6.42 (-9.59, -3.15)	1.94E-04	0.1999
	696.1942	71.6	0.09	9.70 (4.41, 15.25)	2.98E-04	0.1999
	709.994	22	-0.10	-9.62 (-14.33, -4.64)	2.66E-04	0.1999
	879.8333	59.5	0.26	30.15 (14.50, 47.95)	7.43E-05	0.1999
1 <sup>st</sup> trimester	<b>391.2842</b>	<b>22</b>	0.11	11.71 (6.15, 17.56)	3.07E-05	0.1731
	<b>392.2877</b>	<b>22</b>	0.27	31.41 (15.41, 49.63)	5.22E-05	0.1960
	459.3015	22.1	0.11	12.00 (6.30, 18.01)	3.07E-05	0.1731
	94.521	281.7	-0.04	-3.59 (-5.04, -2.13)	3.70E-06	0.0104
	104.0087	294.4	-0.07	-6.63 (-9.31, -3.86)	7.00E-06	0.0157
	143.5513	62.6	-0.06	-5.36 (-7.18, -3.51)	1.00E-07	0.0009
	152.0221	38.1	-0.05	-4.97 (-7.21, -2.67)	4.17E-05	0.0568
	170.9852	55.8	-0.02	-1.89 (-2.79, -0.98)	6.72E-05	0.0689
	188.0231	280.9	-0.05	-4.44 (-6.54, -2.29)	8.71E-05	0.0745
	268.1082	81.9	-0.04	-3.80 (-5.70, -1.86)	1.81E-04	0.1321
One-week	282.1185	119.1	-0.07	-6.46 (-9.27, -3.56)	2.69E-05	0.0432
	285.0238	144.7	-0.08	-7.36 (-10.01, -4.63)	6.00E-07	0.0032
	289.936	65.6	0.03	3.07 (1.47, 4.69)	1.88E-04	0.1321
	342.1238	84.3	0.06	6.02 (3.02, 9.12)	9.25E-05	0.0745
	431.6269	42.5	-0.04	-4.20 (-5.87, -2.49)	3.30E-06	0.0104
	432.8308	81.3	-0.02	-2.25 (-3.34, -1.15)	9.14E-05	0.0745

	454.8629	53.6	-0.04	-3.77 (-5.43, -2.09)	2.13E-05	0.0399
	727.9329	169.7	-0.04	-3.65 (-5.50, -1.76)	2.26E-04	0.1499
	741.9338	59.7	-0.05	-4.55 (-6.66, -2.39)	6.42E-05	0.0689
	969.3444	51.4	-0.06	-6.20 (-8.98, -3.33)	4.53E-05	0.0568
One-month	<b>391.2842</b>	<b>22</b>	0.08	7.95 (4.51, 11.50)	6.20E-06	0.0693

Abbreviations: HILIC, hydrophilic interaction liquid chromatography; mz, mass-to-charge ratio;

RT, retention time; CI, confidence interval;  $FDR_{B-H}$ , Benjamini-Hochberg adjusted  $p$ -values.

**Table S2.** Model statistics of metabolic features in the C18 column associated ( $FDR_{B-H} < 0.2$ ) with ambient  $PM_{2.5}$  exposures of different time windows among 329 pregnant people in the Atlanta African American Maternal-Child Cohort, 2014-2018.

Exposure	mz	RT	Coefficients <sup>a</sup>	Percent change <sup>b</sup> (95% CI)	p-values	FDR <sub>B-H</sub>
1 <sup>st</sup> trimester	473.2826	215	-0.04	-4.23 (-6.02, -2.40)	1.22E-05	0.1166
	529.3236	38.8	-0.15	-13.51 (-19.14, -7.49)	3.38E-05	0.1618
One-week	145.0698	176.9	-0.05	-4.80 (-6.59, -2.99)	7.00E-07	0.0017
	148.0516	284.1	-0.01	-1.29 (-1.70, -0.87)	4.75E-09	<0.0001
	180.9729	25.6	-0.01	-1.38 (-1.78, -0.98)	1.76E-10	<0.0001
	198.0707	28.6	-0.03	-3.23 (-4.95, -1.48)	3.98E-04	0.1963
	288.7788	286.5	-0.05	-4.47 (-6.43, -2.46)	2.39E-05	0.0254
	292.0165	18.7	-0.04	-4.28 (-6.54, -1.96)	4.11E-04	0.1963
	309.1906	210.4	-0.07	-6.56 (-9.35, -3.67)	1.87E-05	0.0254
	343.1673	27.2	-0.05	-4.48 (-5.91, -3.02)	1.21E-08	<0.0001
	367.0712	23.1	-0.08	-7.87 (-11.67, -3.90)	1.80E-04	0.1145
	396.3118	209	-0.04	-3.43 (-5.20, -1.63)	2.76E-04	0.1554
	398.2733	215.1	-0.07	-6.52 (-9.34, -3.62)	2.28E-05	0.0254
	449.2912	270.4	-0.05	-4.88 (-7.04, -2.67)	2.86E-05	0.0273
	458.9445	26.2	-0.04	-3.71 (-5.55, -1.83)	1.69E-04	0.1145
	564.7968	127.2	-0.05	-4.69 (-6.74, -2.60)	2.16E-05	0.0254
	608.2962	234.2	-0.05	-4.80 (-7.29, -2.24)	3.42E-04	0.1818
	639.5576	297.5	-0.04	-3.55 (-5.34, -1.72)	2.03E-04	0.1215
	729.9286	38.8	-0.03	-3.12 (-4.68, -1.53)	1.72E-04	0.1145
	830.5955	268.6	-0.03	-3.35 (-5.03, -1.65)	1.73E-04	0.1145
	933.5317	236.5	-0.04	-3.93 (-5.71, -2.13)	3.61E-05	0.0314
	1036.7953	274.7	-0.07	-6.88 (-9.56, -4.11)	3.20E-06	0.0062
One-month	666.7863	158.4	0.05	5.12 (2.79, 7.50)	1.94E-05	0.1852

Abbreviations: C18, hydrophobic reversed-phase chromatography; mz, mass-to-charge ratio;

RT, retention time; CI, confidence interval; FDR<sub>B-H</sub>, Benjamini-Hochberg adjusted *p*-values.

**Table S3.** Number of overlapping metabolic features associated with PM<sub>2.5</sub> exposures and preterm or early-term birth under different significance cutoff among 329 pregnant people in the Atlanta African American Maternal-Child Cohort, 2014-2018.

Exposure	Outcome	HILIC column			C18 column		
		FDR <sub>B-H</sub> < 0.2	Top 100 <sup>a</sup>	Raw $p < 0.05$	FDR <sub>B-H</sub> < 0.2	Top 100 <sup>a</sup>	Raw $p < 0.05$
One-year	Preterm	0	0	38	0	1	25
	Early-term	0	1	99	0	3	53
1 <sup>st</sup> trimester	Preterm	0	0	29	0	0	23
	Early-term	0	2	66	0	0	34
One-week	Preterm	0	0	25	0	0	17
	Early-term	0	0	18	0	1	13
One-month	Preterm	0	0	24	0	0	24
	Early-term	0	3	41	0	0	26

Abbreviations: PM<sub>2.5</sub>, fine particulate matter; HILIC, hydrophilic interaction liquid chromatography; C18, hydrophobic reversed-phase chromatography; FDR<sub>B-H</sub>, Benjamini-Hochberg adjusted  $p$ -values.

<sup>a</sup> Top 100 metabolic features sorted by raw  $p$ -values.

**Table S4.** Chemical identity with level 1 confidence of the overlapping metabolic features detected at raw *p*-values < 0.05.

mz	RT	Identity	Adduct form	Association	KEGG Pathway <sup>a</sup>
A. HILIC column					
189.1636	21.5	N6,N6,N6-Trimethyl-L-lysine	M+H	1 <sup>st</sup> trimester ~ PTB	Lysine degradation
244.0790	51.8	N-Acetyl-D-Galactosamine	M+Na	1 <sup>st</sup> trimester ~ PTB; One-week ~ PTB; One-month ~ PTB	Galactose metabolism
347.2215	24.5	Cortexolone	M+H	One-week ~ PTB; One-week ~ ETB	Cortisol synthesis and secretion
124.0412	119.9	Picolinic acid	M+H	One-year ~ ETB	Tryptophan metabolism
176.1030	73.0	Citrulline	M+H	One-year ~ ETB	Arginine biosynthesis
177.1063	72.8	Serotonin	M+H	One-year ~ ETB	Tryptophan metabolism
211.0964	26.7	3-Methoxy-L-Tyrosine	M+H	One-year ~ ETB	-
B. C18 column					
162.0196	26.6	N-Acetyl-L-Cysteine	M-H	One-year ~ PTB	-
157.1233	258.2	Nonanoate	M-H	1 <sup>st</sup> trimester ~ PTB; One-week ~ PTB; One-month ~ PTB	-
174.0885	20.9	Citrulline	M-H	One-year ~ ETB	Arginine biosynthesis

Abbreviations: mz, mass-to-charge ratio; RT, retention time; HILIC, hydrophilic interaction liquid chromatography; C18, hydrophobic reversed-phase chromatography.

<sup>a</sup> The biological pathway to which the identified metabolite belongs in human body according to KEGG database.

## **Chapter 5 Integration of the Metabolome and Epigenome in Unravelling the Biological Mechanisms Underlying the Relationship between Ambient PM<sub>2.5</sub> and Preterm and Early-term Birth**

### **Introduction**

Fine particulate matter (PM<sub>2.5</sub>) is a serious threat to public health and has been linked to adverse effects on people of all ages.<sup>132,133,190</sup> Pregnant people and their fetuses are particularly vulnerable and susceptible to the risks caused by PM<sub>2.5</sub> exposure.<sup>10</sup> Previous research has shown that exposure to PM<sub>2.5</sub> during pregnancy is associated with various unfavorable birth outcomes, such as preterm birth (PTB) and early-term birth (ETB).<sup>10</sup> PTB (deliveries before 37 weeks of pregnancy) and ETB (deliveries between 37-39 weeks of pregnancy),<sup>134</sup> are major causes of neonatal illness and death,<sup>135</sup> with adverse effects on mothers<sup>191</sup> and lifelong risks to the children.<sup>136</sup> It is estimated that 10% of PTB cases worldwide can be attributed to PM<sub>2.5</sub> exposure, with the highest impact in sub-Saharan Africa.<sup>192</sup> In the United States, communities of color and those at low socioeconomic status, particularly African Americans (AA), have a higher risk of PTB and ETB.<sup>138,139</sup> Despite a significant amount of research on the association between prenatal PM<sub>2.5</sub> exposure and PTB/ETB, the findings are limited for the African American population.<sup>193</sup> To reduce the incidence of PTB and ETB among vulnerable groups, it is crucial to understand how PM<sub>2.5</sub> exposure leads to health disparities in utero and amplified postnatally, which will inform policies and interventions aimed at prevention.

Omics techniques are a set of techniques used to analyze biological molecules at different molecular levels, including gene (genomics), DNA methylation (epigenomics), mRNA (transcriptomics), and metabolite (metabolomics), which enable a detailed investigation into the biological mechanisms underlying human phenotypes.<sup>194</sup> These techniques are increasingly being used in environmental health sciences to study the complex interplay between

environmental contaminants and health outcomes.<sup>195</sup> For instance, the use of high-throughput metabolomics has been increasingly prevalent in environmental health research for identifying the biological perturbations that result from the long-term and short-term exposures to air pollution.<sup>17</sup> Likewise, researchers have employed metabolomics to identify metabolites and pathways predictive of PTB,<sup>144</sup> and part of the biomarkers were also reported to be altered upon exposure to air pollution among pregnant women, which have important roles in multiple biological pathways, including amino acid metabolism, lipid metabolism, and inflammation.<sup>27,141,145,146,196</sup> The association of DNA methylation (DNAm) in cord blood and placental tissues with prenatal exposure to air pollution has been extensively studied,<sup>197-200</sup> while relatively underexplored in whole blood. DNAm in various genes, including insulin-like growth factor 2 (*IGF2*),<sup>199,201</sup> caspase-7 (*CASP7*),<sup>197</sup> BH3 interacting-domain death agonist (*BID*),<sup>201</sup> serum paraoxonase 1 (*PON1*),<sup>198</sup> paraoxonase 3 (*PON3*),<sup>198</sup> cluster of differentiation 6 (*CD6*),<sup>198</sup> inactive dipeptidyl peptidase 10 (*DPP10*),<sup>202</sup> and 11 $\beta$ -hydroxysteroid dehydrogenase 2 (*HSD11B2*),<sup>203</sup> was found to be associated with air pollutants during different time periods of pregnancy, and the detected genes were found to participate in fetal growth (*IGF2*),<sup>199,201</sup> apoptosis (*CASP7* and *BID*),<sup>197,201</sup> lipid metabolism (*PON1* and *PON3*),<sup>198</sup> inflammation (*CD6* and *DPP10*),<sup>198,202</sup> and steroid hormone metabolism (*HSD11B2*).<sup>203</sup> DNAm has also been investigated as a major epigenetic mechanism for PTB.<sup>204</sup> The associations of cg03915055 [promoter region of cytohesin 1 interacting protein (*CYTIP*)] and cg06804705 [promoter region of long intergenic non-protein coding RNA 114 (*LINC00114*)] with spontaneous PTB were reported among 150 pairs of African American mother-newborn pairs in Boston.<sup>205</sup>

The rapid generation of omics data further motivates researchers to integrate multiple omics data to obtain a more comprehensive view of disease development.<sup>16</sup> The integration of multi-omics data can contribute to building evidence on causation by identifying the biomarkers indicative of changes at another molecular level and cross-validating the functional activity of identified biological pathways. For example, by integrating genome-wide DNAm and metabolomic



data, we can enhance our understanding of pollutant-associated differences in metabolic profiles through the annotation of related CpG sites. A few recent studies based on this integrative approach have highlighted the interplay between DNAm and metabolomic changes, which proposed key mechanistic pathways related to environmental contaminants.<sup>206</sup> A joint metabolomic-epigenomic study was conducted to investigate the relationship between smoking-related metabolites and DNAm.<sup>207</sup> Three endogenous plasma metabolites - N-Acetylpyrrolidine, 5-Hydroxycotinine, and 8-Oxoguanine - were found to be associated with 3-16 CpG sites. These sites were located in genes involved in cell proliferation and apoptosis evasion, which are known to play a role in the development of cancer.<sup>207</sup> Short-term exposure to ambient ozone was found to be associated with an increased level of circulating angiotensin-converting enzyme (ACE) and endothelin-1 (ET-1) among a panel of healthy college students. In the meanwhile, a decreased level of DNAm of *ACE* and *ET-1* genes was also found.<sup>208</sup> ACE and ET-1 are important factors regulating blood pressure, and an association between ozone exposure and elevated blood pressure was observed in the study.<sup>208</sup>

Given the aforementioned evidence on the potential interplay of metabolic profiles and DNAm, it is logically to hypothesize that the integration of metabolomic and genome-wide DNAm can reveal biological pathways mediating the adverse effect of PM<sub>2.5</sub> on PTB and ETB. However, to the best of our knowledge, no study has been conducted to explore this hypothesis. To address this critical knowledge gap, we conducted the current study in the prospective Atlanta African American (ATL AA) Maternal-Child Cohort, which was established in 2014 to assess the impact of the prenatal environmental contaminants on adverse birth outcomes.<sup>147</sup> We aimed to employ multi-omics integration techniques to identify changes in metabolic profiles and DNAm and assess their interrelationship and how they mediate the association between short- and long-term exposures to ambient PM<sub>2.5</sub> and the risk of PTB and ETB.

## Methods

### *Study population*

The study involved individuals who were recruited in the ATL AA Maternal-Child Cohort.<sup>147,148</sup>

This prospective cohort has been enrolling pregnant people who self-report as U.S.-born African American, are between the ages of 18 and 40, have singleton pregnancies, and no chronic medical conditions, during their first prenatal care visits between 8 and 14 weeks of gestation, at Emory Midtown Hospital (privately funded) and Grady Hospital (publicly funded) since 2014. Data was collected using questionnaires, medical record abstraction, and blood samples taken during routine blood draws. The participants were tracked until delivery to obtain information on birth outcomes. The serum samples and peripheral blood mononuclear cells (PBMC) collected during their first prenatal care visit were stored at -80°C and were analyzed using untargeted metabolomics analysis and genome-wide DNAm analysis respectively, using established protocols.<sup>147</sup> A total of 264 AA pregnant people with valid metabolomics and DNAm data available at their first visit were included in the study, all of whom were enrolled between May 2014 and May 2018. The Emory University Internal Review Board approved this study, and all participants provided signed informed consent.

### *Air pollution exposure assessment*

To estimate individual exposure, we used a well-established ensemble-based model with a spatial resolution of 1km to spatially join the geocoded residential addresses provided during the first visit from each participant and obtain daily PM<sub>2.5</sub> concentrations. This model integrated multiple machine learning algorithms and predictor variables, and its specifics were previously published.<sup>149</sup> To ensure that the exposure preceded the mediator (i.e., metabolomics and DNAm data), we selected four averaging periods: 1-year prior to conception, the first trimester, 1-month

prior to blood draw, and 1-week prior to blood draw, which helped distinguish between long-term and short-term exposures.

### *Measure of birth outcomes*

To determine whether a participant had a preterm birth (PTB) or early term birth (ETB), we relied on the gestational age at delivery, which we calculated using the delivery date in relation to the estimated date of confinement, as per the guidelines of the American College of Obstetrics and Gynecology (ACOG),<sup>145</sup> based on information abstracted from medical records. The estimated date of confinement was established for all participants at their first prenatal care visit using the last menstrual period and/or first trimester ultrasound.<sup>150</sup> We categorized the participants into one of three groups: PTB (live birth occurring after 20 weeks but before 37 weeks), ETB (live birth occurring at 37 weeks or later but before 39 weeks), and full-term birth (FTB, birth occurring at 39 weeks or later).<sup>134</sup> In other words, two binary outcomes variables were created both with the FTB group as the reference.

### *High-resolution metabolomics*

We used high-resolution liquid chromatography coupled with mass spectrometry (HR-LCMS, Thermo Scientific™ Q- Exactive™ HF) to analyze non-fasting samples in triplicate. The details of the established protocol can be found elsewhere<sup>147,151</sup>. Briefly, we employed two analytical columns, hydrophilic interaction liquid chromatography (HILIC) with positive electrospray ionization (ESI) and C18 hydrophobic reversed-phase chromatography with negative ESI, to detect metabolomic profiles. The metabolic features were extracted from the profiles by R packages *apLCMS* with *xMSanalyzer* with mass-to-charge ratio (mz), retention time (rt), and relative intensity.<sup>152,153</sup> Then, the relative intensity for each sample was averaged across triplicates and transformed with the natural log for down-stream analysis. The relative standard deviation (RSD) was calculated for each feature across quality control (QC) samples, and

missingness was estimated for both study and QC samples. We excluded features with RSD > 50% and QC missingness < 10%, or study sample missingness < 90%, resulting in 11,269 and 9,565 metabolic features for the HILIC and C18 columns, respectively. To impute missing values, we used either quantile regression imputation of left-censored data (QRILC) or random forest (RF), depending on the missing pattern.<sup>154</sup> We classified the missing pattern [missing not at random (MNAR) vs. missing at random (MAR)] using a second, correlated auxiliary feature with Pearson's correlation > 0.5, and assumed that insights into the pattern of missing values could be gained from corresponding non-missing observations of the auxiliary feature.<sup>155</sup> We used QRILC to impute missing values of MNAR features and RF to impute missing values of MAR features.

### *Genome-wide DNA methylation*

Methylation data for the participants were generated on the Illumina Infinium MethylationEPIC BeadChip. All DNAm data were preprocessed to remove low-quality samples and probes through the use of the R package *minfi*.<sup>87</sup> Raw intensity files were converted to a data frame consisting beta values [Methylated signal/(Methylated signal + Unmethylated signal)] ranging on a continuous scale from 0 to 1 for each of the CpG sites. Samples with more than 1% probes having a detection p value over 1% were excluded. Probes with more than 1% of samples having a detection p value over 1% were excluded. Additionally, cross-reactive probes predicted were also excluded.<sup>209</sup> Then, the Combat algorithm was used to normalize the distributions to reduce technical variation and correct for differences between type I and type II probe signals.<sup>85</sup> In total, after all preprocessing steps, 839,487 CpG sites remained for the down-stream analysis. We used the R package *EpiDISH* to obtain the cell-type proportions (CD4+ T cells, CD8+ T cells, natural killer cells, B cells, and monocytes) in the PBMC using the reference dataset published previously.<sup>210,211</sup>

### *Covariate assessment*

We drew the confounding structure from a literature review and our previous studies.<sup>145,151,212</sup>

The resulting confounding structure was visualized through directed acyclic graphs (DAGs), which can be found in the Supplement (Figure S1). We collected individual-level demographic characteristics, including maternal age and maternal educational attainment (classified as less than high school, high school, or some college or more), via a standardized interview questionnaire. Information on infant sex (binary), parity (classified as nulliparity, primiparity, or multiparity), and tobacco and marijuana use in the month before pregnancy (binary), as well as alcohol use in the month prior to pregnancy (binary), were abstracted from medical records. Maternal body mass index (BMI, kg/m<sup>2</sup>) was calculated using measurements of weight and height taken at the first visit. To control for meteorological factors, we included the conception season (for long-term exposure) and averaged apparent temperature (for short-term exposure within the same time windows as air pollution estimates) as covariates. We obtained the daily apparent temperature at Hartsfield-Jackson International Airport using the R package *riem*.<sup>156</sup> The proportions of four cell types (CD8+ T cells, natural killer cells, B cells, and monocytes) in PBMC were included in the epigenome-wide association analysis to control for the confounding from the cell type composition. We included the same covariates for the metabolomic and epigenomic analyses, except for cell type composition which was only included in epigenomic analysis.

### *Statistical analysis*

To identify the metabolic features and differentially methylated CpG sites associated with both PM<sub>2.5</sub> exposures and birth outcomes, we first subset the high-dimensional metabolomic and genome-wide DNAm data based on the strength of their associations with PM<sub>2.5</sub> exposure, as a means of dimensional reduction. Specifically, we conducted a metabolome-wide association

study (MWAS) and an epigenome-wide association study (EWAS) for  $PM_{2.5}$  exposures using the following equations:

$$\ln(Feature_j) = \beta_{0j} + \beta_{1j}PM_{2.5} + \beta_{2j}Age + \beta_{3j}Education + \beta_{4j}Sex + \beta_{5j}BMI + \beta_{6j}MET + \beta_{7j}Toba\_mari + \beta_{8j}Alcohol + \epsilon_j \quad \text{Eq. (1)}$$

$$Beta_k = \beta_{0k} + \beta_{1k}PM_{2.5} + \beta_{2k}Age + \beta_{3k}Education + \beta_{4k}Sex + \beta_{5k}BMI + \beta_{6k}MET + \beta_{7k}Toba\_mari + \beta_{8k}Alcohol + \beta_{9-12k}celltype + \epsilon_j \quad \text{Eq. (2)}$$

where  $\ln(Feature)$  refers to the natural log of intensity of metabolic feature  $j$ ;  $PM_{2.5}$  is the averaged  $PM_{2.5}$  exposure for a specific window, and  $MET$  is the corresponding meteorological covariate;  $Beta_k$  refers to the beta value of CpG site  $k$ ;  $celltype$  denotes four types of cells in PBMC, including CD8+ T cells, natural killer cells, B cells, and monocytes, while CD4+ T cells served as reference. The R package *CpGassoc* was used to perform EWAS based on Eq. (2).<sup>213</sup> We used Benjamini-Hochberg (BH) approach ( $FDR_{B-H}$ ) and Bonferroni approach (Bon) to account for multiple testing for MWAS and EWAS, respectively.<sup>158,214</sup> Results were presented using Manhattan plots and QQ plots (Figure S2 & S3 in the Supplement).

A joint metabolome-epigenome approach with a focus on exposure-related metabolic features was considered in previous literature to reveal additional details in molecular responses to environmental contaminants.<sup>207</sup> To understand the relationship between DNA methylation and the exposure-related metabolic features, we used the R package *mixOmics* to conduct a regularized canonical correlation analysis (RCCA).<sup>215</sup> RCCA seeks to identify linear relationships between two sets of variables, which is an extension of canonical correlation analysis that allows for the inclusion of regularization penalties to handle high-dimensional data.<sup>216</sup> Specifically, RCCA takes the inputs of metabolomic and epigenomic data and searches for linear combinations of the inputs in each dataset that maximize the correlation between the two datasets.<sup>216</sup> In other words, RCCA can reveal global correlation patterns between metabolic features and CpG sites. In the current analysis, for each  $PM_{2.5}$  exposure window, we included

the significant metabolic features ( $\text{FDR}_{\text{B-H}} < 0.2$ ) and the 100 CpG sites that had the highest pairwise Pearson's correlation with each feature as inputs based on all CpG sites. A relevance network was generated by *mixOmics* with nodes representing correlated features and CpG sites and edges representing correlations (a correlation score  $> 0.5$ ).<sup>215</sup> The strength of the correlations, for absolute score, was regarded as weak for  $< 0.4$ , moderate for  $0.40-0.59$ , strong for  $0.6-0.79$ , and very strong for  $\geq 0.8$ .

To further illustrate the potential mediating role of the interplay between metabolome and genome-wide DNA methylation for the association of  $\text{PM}_{2.5}$  exposure with PTB and ETB, we used the R package *MOFA2* to perform an unsupervised multi-omics integration and conducted causal mediation analysis to identify mediating orthogonal factors derived from *MOFA2*.<sup>217,218</sup> *MOFA2* refers to multi-omics factor analysis and as indicated by its name, is a dimension reduction technique that is able to identify latent factors representing the driving sources of variation across multiple sets of variables.<sup>217</sup> The key idea behind *MOFA2* is that the same latent factors can explain the variation in multiple omics datasets, allowing the integration of information across different modalities. In the current analysis, to fit *MOFA2* into a mediation framework, we first filtered the metabolic features and CpG sites based on their association with  $\text{PM}_{2.5}$  exposure. For each exposure window, the most significant metabolic features and CpG sites from the MWAS and EWAS aforementioned were selected as inputs for *MOFA2*. To lower the type II error and include potential mediators exclusively, we explore two different thresholds of  $p$  (the number of input metabolic features or CpG sites: 500 and 1000). Then, *MOFA2* generated latent factors of an initial number (we specified the parameter as 30) and dropped factors that explain no variance. We extracted the latent factors from the resulting *MOFA2* model for the downstream causal mediation analysis. We used the R package *mediation* to conduct the single mediation analysis with the adjustment of preselected confounding,<sup>218</sup> and the confounding structure for the exposure-mediator and mediator-outcome association was illustrated in the Supplement (Figure S1). In addition, we also employed *mixOmics* to identify

exposure-related metabolic features and CpG sites based on their prediction performance on the birth outcomes (i.e., PTB vs. ETB vs. FTB). Specifically, the same thresholds of  $p$  (the number of input) were used, and the sparse partial least squares-discriminant analysis was conducted to select influential metabolic features and CpG sites.<sup>215</sup>

Separate analyses were conducted for each column (HILIC positive ESI and C18 negative ESI). Then, the metabolic features associated with PM<sub>2.5</sub> exposure or enriched in the latent factors were identified and annotated for functional interpretation, and the differentially methylated CpG sites correlated to the significant metabolic features or enriched in the latent factors were annotated using an online annotation data `IlluminaHumanMethylationEPIC`.<sup>219</sup> Additional functional insight on metabolic features and CpG sites was obtained by search the corresponding analyte in publicly available databases, including KEGG and EWAS catalog.<sup>101,220</sup>

All analyses were completed in R (version 4.2.0).

## Results

### *Study population characteristics*

A total of 264 pregnant people was included in the current analysis, and their characteristics are described in Table 1. The average maternal age was 25.1 years (SD=4.69-4.92) for the PTB, ETB, and FTB groups. The PTB group had a slightly lower average BMI (28.0 kg/m<sup>2</sup>), highest proportion of multiparity (40.4%), highest male-to-female ratio (~3:2), highest proportion of tobacco or marijuana use (50.0%). Half of the participants in the FTB group completed a college degree or higher, which is higher than the other two groups but lower than the number (62.6%) in the U.S. female population of all races over 25 years old.<sup>221</sup> The FTB group had the highest proportion of nulliparity. The medians of one-year, 1<sup>st</sup> trimester, one-week and one-month PM<sub>2.5</sub> exposures were 10.35 [interquartile range (IQR)=1.18], 10.00 (IQR=2.65), 9.91 (IQR=3.63),



9.68 (IQR=3.33)  $\mu\text{g}/\text{m}^3$ , respectively (Table 2). The IQR was slightly higher for the short-term exposures, indicating a higher variation contributed by temporal variability.

#### *Metabolome-wide and epigenome-wide association analysis*

In MWAS, there were 3, 10, 37, and 3 metabolic features in the HILIC column associated ( $\text{FDR}_{\text{B-H}} < 0.2$ ) with  $\text{PM}_{2.5}$  exposure for 1 year prior to conception, the 1<sup>st</sup> trimester, 1 week, and 1 month prior to blood draw, respectively (Table S2). Only 3 and 20 features were found to be associated ( $\text{FDR}_{\text{B-H}} < 0.2$ ) with the 1<sup>st</sup> trimester, 1 week exposures, respectively, in the C18 column. The number of significant metabolic features associated  $\text{PM}_{2.5}$  exposures for each column at different thresholds of significance were summarized in the Supplement (Table S2). One feature (mz: 391.2842, rt: 22) associated with both 1<sup>st</sup>-trimester and one-month exposures was annotated as Di(2-ethylhexyl)phthalate (DEHP) which is a frequently used plasticizer and has a profound impact on human health.<sup>164</sup> One features (mz: 282.1185, rt: 119.1) associated with one-week exposure was annotated as methyladenosine with level 1 confidence. Methyladenosine is a modified form of adenosine where a methyl group is added and is the most common modifications found in mRNA.<sup>163</sup>

EWAS yield all lambda values (i.e. inflation factor) between 0.90 and 1.42 (Figure S3). We did not observe any differentially methylated CpG sites associated with one-year, 1<sup>st</sup> trimester or one-month  $\text{PM}_{2.5}$  exposures after multiple testing correction ( $\text{FDR}_{\text{Bon}} < 0.2$ ), while there were 243 differentially methylated CpG sites for one-week exposure (Table S2).

#### *Metabolome-epigenome network analysis*

Using *mixOmics*, we visualized the correlation patterns between  $\text{PM}_{2.5}$ -associated metabolic features and CpG sites with correlation less than 0.5 excluded (Figure 1). Although we were not able to annotate the metabolic features that had a moderate correlation with at least one CpG site, the correlated CpG sites were matched to corresponding genes according to the online

reference database. One metabolic feature (mz: 152.0221, rt: 38.1) associated with one-week exposure prior to blood draw in the HILIC column was moderately correlated with cg17254259 (intergenic); The other feature (mz: 657.7080, rt: 55.5) in the HILIC column was moderately correlated with 5 (cg00804091, *EMC8*; cg15894315, *DHCR24*; cg13334650, *APBB1*; cg07147709, *MIR7-2*; cg04439028, *TRAP1*) CpG sites; One features (mz: 569.5778, rt: 21.2) associated with 1<sup>st</sup> trimester exposure in the C18 column was strongly correlated with cg08189615 (*PNKD*); The other feature (mz: 529.3236, rt: 38.8) in the C18 column associated with 1<sup>st</sup> trimester was strongly correlated with cg11553068 (*BRF1*); One feature (mz: 564.7968, rt: 127.2) associated with one-week exposure in the C18 column was moderately correlated with 5 (cg18739834, intergenic; cg21384857, intergenic; cg25307371, *LEPROT*; cg07109124, *ARSB*; cg09762612, *FIZ1*); The other feature (mz: 729.9286, rt: 38.8) in the C18 column associated with one-week exposure was moderately negatively correlated with cg27385053 (*WDTC1*).

### *Multi-omics mediation analysis*

The latent factors were generated by *MOFA2*, and those with less than 1% variance explained were dropped. To explore the impact of the number of inputs on the resulting low-dimensional representation, we employed two levels of inputs (500 and 1000, and inputs referring to features and CpG sites) with the consideration of including all potential mediating metabolic features or CpG sites. The factors with indirect effects were summarized in the Supplement, and those with significant indirect effects were showed in Table 3. In total, seven latent factors were found to significantly mediate the associations of PM<sub>2.5</sub> exposures with ETB, while none were found for PTB. Two of the seven factors significantly mediated the positive association of 1<sup>st</sup>-trimester exposure and one-month exposure prior to blood draw, respectively, with ETB. Four out of the seven significant factors were detected at the level of 500 inputs. The associations for one-year exposure prior to conception and one-week exposure to blood draw were consistently

significantly mediated by a latent factor across the two technical columns of metabolome. Of note, the variance explained by the seven factors was mostly contributed by metabolic features (as illustrated by the 6<sup>th</sup> and 7<sup>th</sup> columns in Table 3). To capture an interpretable view of the low-dimensional representation, the top 10 metabolic features and CpG sites with highest weights for each the seven significant factors were illustrated in Figure 2. The consistencies were primarily observed at the level of 1000 inputs. One metabolic feature (mz: 852.0424, rt: 274.8) was consistently detected among the top 10 of the three significant factors at the level of 1000 inputs. One CpG site (cg16401529, *PABPC5*) was consistently found in the significant factors for 1<sup>st</sup>-trimester and one-month exposures. Methyladenosine was among the top 10 metabolic features in the latent factor statistically mediating the association between one-week exposure and ETB.

## Discussion

In the Atlanta African American Maternal Child cohort, we combined metabolomic and genome-wide DNAm data to identify metabolic features and CpG sites that potentially mediated the association between prenatal PM<sub>2.5</sub> exposure with PTB and ETB. Using a metabolome-epigenome network analysis, we identified six PM<sub>2.5</sub>-related metabolic features that were moderately or strongly associated with 1-5 CpG sites, indicative of the interplay of metabolome and epigenome underlying the biological responses of pregnant people to PM<sub>2.5</sub> exposure. By combining multi-omics factor analysis and causal mediation analysis, we identified seven latent factors that, respectively, significantly mediated the associations of PM<sub>2.5</sub> exposure for different windows with ETB. Of note, one metabolic feature (mz: 282.1185, rt: 199.1), annotated as methyladenosine, was found among the top 10 features having highest weights for a latent factor that significantly mediated the association between one-week exposure prior to blood

draw and ETB, which was also found to be statistically associated with the same exposure window by MWAS after multiple comparison correction.

Methyladenosine is the most abundant form of post-transcriptional modification found in mRNA, which is known to serve as a regulatory factor in several critical biological processes, such as RNA splicing, RNA stability, and translation.<sup>222</sup> In the past few years, mounting evidence has emerged highlighting the critical role of methyladenosine in modulating the toxicity of environmental contaminants.<sup>223</sup> A negative dose-response relationship was reported between the global methyladenosine RNA methylation level and the concentration of particulate matter exposure in an *in-vitro* study where the human lung epithelial cells were exposed for a duration of 24-48 hours.<sup>224</sup> In the meanwhile, the expression of methyladenosine ‘eraser’, enzymes removing the methyl group, was found to be upregulated among human subjects exposed to a higher level of PM<sub>2.5</sub>.<sup>224</sup> Among a panel of 120 adults, long-term smoking status, quantified in pack-year, was observed to have a negative association with global methyladenosine RNA methylation in the peripheral blood.<sup>163</sup> Using a mouse model, Li et al. (2019) found that the exposure to PM<sub>2.5</sub> for 24 hours resulted in an increase in the global level of methyladenosine in lung tissues, which was restored to baseline levels following air purification for 120 hours,<sup>225</sup> implying an irreversible effect of PM<sub>2.5</sub> exposure. In the current analysis, we also found two CpG sites with high weights for the same significant factor as methyladenosine, but they are in the intergenic region. More investigations are warranted to elucidate the function of RNA methylation in the PM<sub>2.5</sub>-induced ETB.

The metabolome-epigenome network analysis identified six PM<sub>2.5</sub>-related metabolic features that were moderately or strongly associated with 1-5 CpG sites. To our best knowledge, none of these CpG sites have been reported in previous literature to be associated with air pollution exposure. One CpG site (cg15894315), in 24-Dehydrocholesterol reductase (*DHCR24*) gene, was moderately correlated with a metabolic feature associated with one-week PM<sub>2.5</sub> exposure. *DHCR24* encodes an enzyme participating in the biosynthesis of cholesterol, and its RNA

transcript was suggested in a previous study to serve as a potential transcriptional biomarker of air pollution effect.<sup>226</sup> Another CpG site (cg04439028), in TNF receptor associated protein (*TRAP1*) gene, was associated with the same metabolic feature. *TRAP1* is a mitochondrial protein involved in cell protection from oxidative stress, and the transcription of *TRAP1* was reported to be modulated after an acute exposure to coarse particulate matter in the human monocyte-macrophage cells.<sup>227</sup> Although the association of DNAm with prenatal exposure to air pollution has been extensively studied, the previous studies were mainly conducted based on cord blood and placental tissues.<sup>197-200</sup> Moreover, the role of DNAm in the air-pollution-associated perturbations in the metabolic profiles was furtherly unexplored as far as we know.

We for the first time used the multi-omics integration approach to identify the metabolic features and CpG sites that potentially mediated the association between PM<sub>2.5</sub> exposure and ETB. However, the application of multi-omics integration is still nascent and remains challenging for use in environmental epidemiology for a few reasons. One of the primary difficulties is the high dimensionality of omics data. Integrating omics data usually requires the use of dimension reduction and regularization methods, while there is no universally accepted workflow.<sup>228</sup> In the current analysis, we employed a two-stage reduction based on our study objective: 1) metabolic features or CpG sites filtered by their associations with the exposure of interest; 2) the integration tools (i.e., *mixOmics* and *MOFA2*) with a built-in regularization algorithm. One caveat of our study is that while our findings suggest a potential interplay of metabolome and epigenome, further studies with larger sample sizes are needed to validate our findings. Additionally, the multi-omics integration can present challenges in terms of data interpretation, as it requires advanced statistical and computational methods to extract meaningful insights from large, complex datasets. We addressed this challenge to some extent by focusing on the exposure-associated metabolic features and leveraging the weights of individual features or CpG sites on the low-dimensional representation.

Our analysis adds to a growing body of research investigating underlying mechanisms of air-pollution-related adverse birth outcome by characterizing maternal metabolomic profiles.<sup>27</sup> Previous studies that examined the associations of the maternal metabolomics perturbations with air pollution exposure found that the levels of serum serine, creatinine, histidine, myo-inositol, heptadecanoic acid, and linoleic acid in mid pregnancy were associated with high traffic-related air pollution during the first trimester among women of multiple races;<sup>141</sup> five steroids in mid- and late-pregnancy were associated with PM<sub>2.5</sub> exposure during early pregnancy in an ethnically diverse population.<sup>171</sup>

Our study has several notable strengths in addition to the novel strategy combining multi-omics integration and mediation analysis. Gestational age was accurately determined during the early pregnancy visit. The analytical cohort we examined comprised exclusively of individuals of African ancestry, a population that has been considerably underrepresented in environmental epidemiological investigations. Lastly, the workflow of the untargeted metabolomics profiling was thoroughly established and has previously demonstrated its efficacy in analyzing numerous non-fasting samples.<sup>146,189</sup>

There are several limitations in the current analysis and we also listed potential avenues for future research. First, our findings cannot provide inference on causal relationship. The perturbations observed in omics data may potentially arise due to the presence of co-exposure or unmeasured confounding, which should be further explored via multipollutant approach. Second, the serum samples were obtained during early and mid-pregnancy, and we were not able to examine other crucial exposure time frames. Third, the estimated ambient PM<sub>2.5</sub> exposure was estimated solely based on the residential address, which did not account for daily mobility patterns and exposures in other microenvironments. Fourth, as indicated by the lambda values of EWAS, the model results for PM<sub>2.5</sub>-DNAm associations were inflated/deflated, suggesting the uncontrolled confounding. Last, the sequential relationship between changes in metabolome and DNA methylome was not assessed in the current analysis. Advanced

statistical techniques and epidemiological designs combined with *in-vivo* models are warranted to disentangle the regulation process.

## Conclusions

Using the multi-omics integration and mediation analysis, we identified latent factors, derived from metabolomic and epigenomic data, that significantly mediated the association between PM<sub>2.5</sub> exposure and the risk of ETB in a marginalized and understudied population. Our findings provide important information about the interplay between metabolic features and DNAm that appears to play a critical role in PM<sub>2.5</sub> toxicity. The integration of metabolome and DNA methylome provides great potential for capturing a more holistic picture of the biological mechanisms underlying PM<sub>2.5</sub> toxicity on pregnancy outcomes and lend support to future development of biomarker.

## Tables and figures

**Table 1.** Selected population characteristics by birth outcomes among the subjects enrolled in Atlanta African American Maternal-Child Cohort study, 2014-2018 (N = 264).

	Preterm (N=52)	Early-term (N=46)	Full-term (N=166)	<i>p</i> -value
Maternal age, mean (SD)	25.1 (4.69)	25.0 (4.79)	25.1 (4.92)	0.984
Maternal educational attainment, No. (%)				
Less than high school	9 (17.3%)	12 (26.1%)	22 (13.3%)	0.081
High school	24 (46.2%)	19 (41.3%)	61 (36.7%)	
Some college or more	19 (36.5%)	15 (32.6%)	83 (50.0%)	
Body mass index, mean (SD)	28.0 (7.09)	28.3 (8.28)	29.3 (7.68)	0.442
Parity, No. (%)				
Nulliparity	19 (36.5%)	14 (30.4%)	80 (48.2%)	0.045
Primiparity	12 (23.1%)	19 (41.3%)	43 (25.9%)	
Multiparity	21 (40.4%)	13 (28.3%)	43 (25.9%)	
Infant sex, No. (%)				
Male	32 (61.5%)	23 (50.0%)	74 (44.6%)	0.096
Female	20 (38.5%)	23 (50.0%)	92 (55.4%)	
Maternal tobacco or marijuana use, No. (%)				
No	26 (50.0%)	28 (60.9%)	89 (53.6%)	0.554
Yes	26 (50.0%)	18 (39.1%)	77 (46.4%)	
Maternal alcohol use, No. (%)				
No	48 (92.3%)	42 (91.3%)	150 (90.4%)	0.951
Yes	4 (7.7%)	4 (8.7%)	16 (9.6%)	
Season of conception, No. (%)				
Spring	14 (26.9%)	10 (21.7%)	38 (22.9%)	0.979
Summer	16 (30.8%)	14 (30.4%)	51 (30.7%)	
Fall	13 (25.0%)	12 (26.1%)	38 (22.9%)	
Winter	9 (17.3%)	10 (21.7%)	39 (23.5%)	
Proportions of cell type, mean (SD)				
CD4+ T cell	0.304 (0.068)	0.298 (0.057)	0.308 (0.067)	0.675
CD8+ T cell	0.238 (0.060)	0.249 (0.048)	0.223 (0.062)	0.019
Natural killer cell	0.089 (0.029)	0.092 (0.032)	0.096 (0.036)	0.326
B cell	0.107 (0.041)	0.119 (0.039)	0.105 (0.035)	0.087



	Preterm (N=52)	Early-term (N=46)	Full-term (N=166)	<i>p</i> -value
Monocyte	0.231 (0.079)	0.213 (0.060)	0.233 (0.083)	0.320

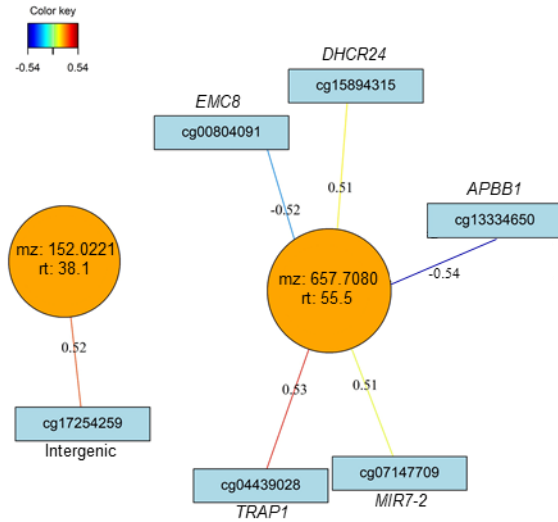
Abbreviations: SD, standard deviation.

**Table 2.** Statistics of fine particulate matter (PM<sub>2.5</sub>) exposure for the four exposure windows among 264 pregnant people in the Atlanta African American Maternal-Child Cohort, 2014-2018.

PM <sub>2.5</sub> exposure (µg/m <sup>3</sup> )	Q1	Q3	IQR	Median	Mean
One-year exposure prior to conception	9.72	10.90	1.18	10.35	10.30
Exposure for the 1 <sup>st</sup> trimester	9.06	11.71	2.65	10.00	10.50
One-week exposure prior to blood draw	8.18	11.82	3.63	9.91	10.39
One-month exposure prior to blood draw	8.46	11.79	3.33	9.88	10.50

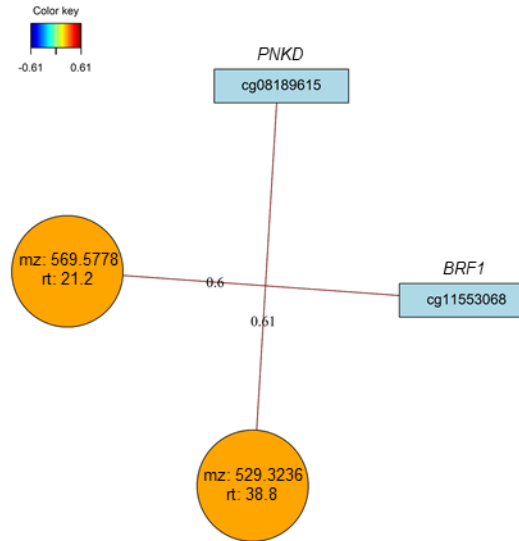
Abbreviations: Q1, 1<sup>st</sup> quartile; Q3, 3<sup>rd</sup> quartile; IQR, interquartile range.

A. HILIC column

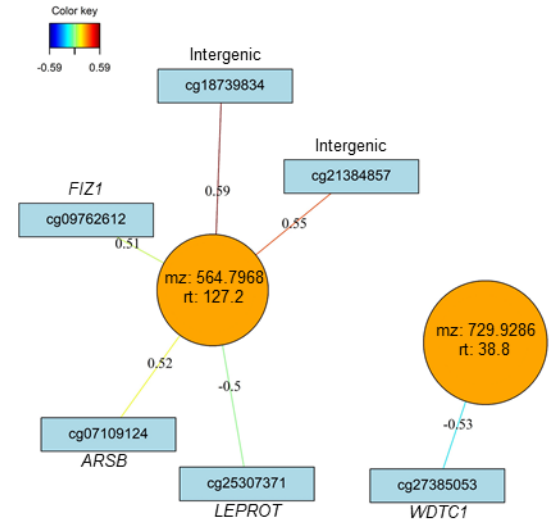


One-week PM<sub>2.5</sub> exposure  
prior to blood draw

B. C18 column



1<sup>st</sup> trimester PM<sub>2.5</sub> exposure



One week PM<sub>2.5</sub> exposure  
prior to blood draw

**Figure 1.** Relevance network showing the correlations among PM<sub>2.5</sub>-associated metabolic features and CpG sites with a correlation threshold of 0.5. The PM<sub>2.5</sub>-associated metabolic features without correlated CpG sites (i.e., correlation score > 0.5) were not showed in the figures. Abbreviations: HILIC, hydrophilic interaction liquid chromatography; C18, hydrophobic reversed-phase chromatography; *mz*, mass-to-charge ratio; *rt*, retention time; PM<sub>2.5</sub>, fine particulate matter.

**Table 3.** Indirect effect estimated by causal mediation analysis via the R package *mediation* of latent factors generated by *MOFA2* for the associations between PM<sub>2.5</sub> exposure and early-term birth <sup>a</sup>.

Data type	$P^b$	Total number of latent factors	Exposure	Factor index <sup>c</sup>	Variance explained		ACME <sup>d</sup>	Total effect <sup>e</sup>
					per omics data			
					Metabolome	DNAm		
HILIC + DNAm	500	2	One-year	1	7.900	4.44×10 <sup>-3</sup>	-0.04 (-0.09, -0.01)	-0.04 (0.05, -0.13)
		2	1 <sup>st</sup> trimester	-	-	-	-	-
		6	One-week	3	2.616	5.04×10 <sup>-5</sup>	-0.01 (-0.04, 0.00)	-0.03 (0.02, -0.12)
		4	One-month	-	-	-	-	-
	1000	8	One-year	4	4.638	2.78×10 <sup>-4</sup>	-0.04 (-0.09, -0.01)	-0.04 (0.04, -0.13)
		4	1 <sup>st</sup> trimester	1	5.953	3.43×10 <sup>-4</sup>	0.04 (0.01, 0.10)	0.00 (0.11, -0.11)
		7	One-week	-	-	-	-	-
		8	One-month	1	3.548	6.44×10 <sup>-5</sup>	0.03 (0.00, 0.10)	0.00 (0.08, -0.10)
C18 + DNAm	500	5	One-year	4	1.543	4.46×10 <sup>-4</sup>	-0.02 (-0.04, 0.00)	-0.04 (0.04, -0.13)
		4	1 <sup>st</sup> trimester	-	-	-	-	-
		4	One-week	4	1.932	3.04×10 <sup>-3</sup>	-0.02 (-0.06, 0.00)	-0.06 (0.02, -0.16)
		3	One-month	-	-	-	-	-
	1000	8	One-year	-	-	-	-	-
		7	1 <sup>st</sup> trimester	-	-	-	-	-
		7	One-week	-	-	-	-	-
		7	One-month	-	-	-	-	-

Abbreviations: PM<sub>2.5</sub>, fine particulate matter; chr, chromosome; ACME, average causal mediated effect (i.e., indirect effect).

<sup>a</sup> The factors with significant indirect effect were only observed for the associations with early-term birth.

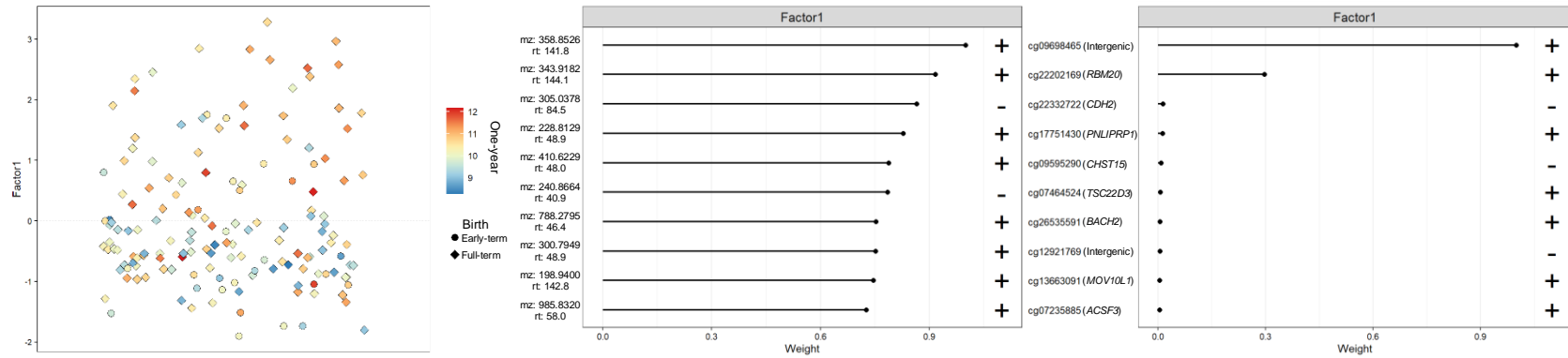
<sup>b</sup> The number of input metabolic features or CpG sites entered the multi-omics factor analysis to generate latent factors.

<sup>c</sup> The factors were indexed in the order of the amount of variance explained.

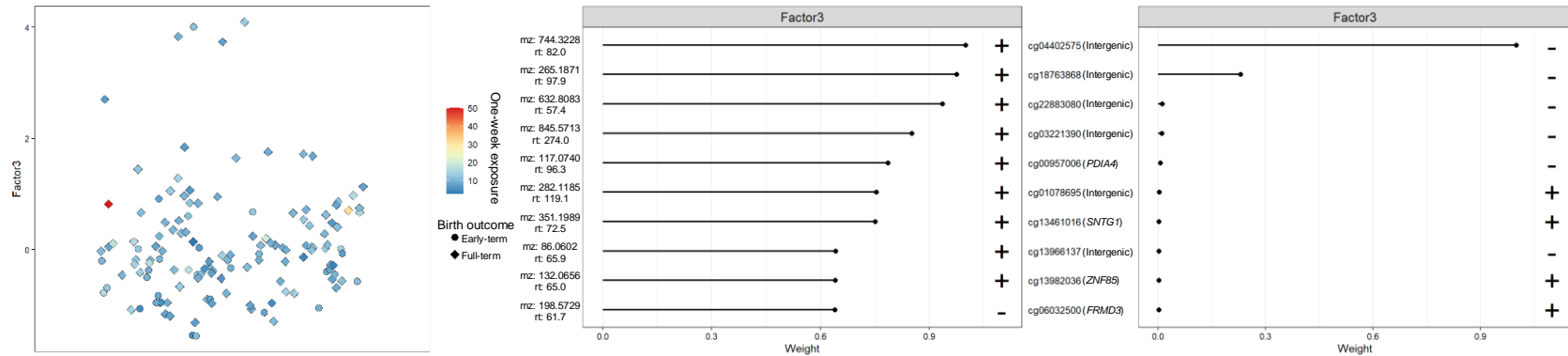
<sup>d</sup> The ACME was estimated regarding one-unit increase in the factor value.

<sup>e</sup> The total effect was estimated regarding one-interquartile-range increase in PM<sub>2.5</sub> exposures.

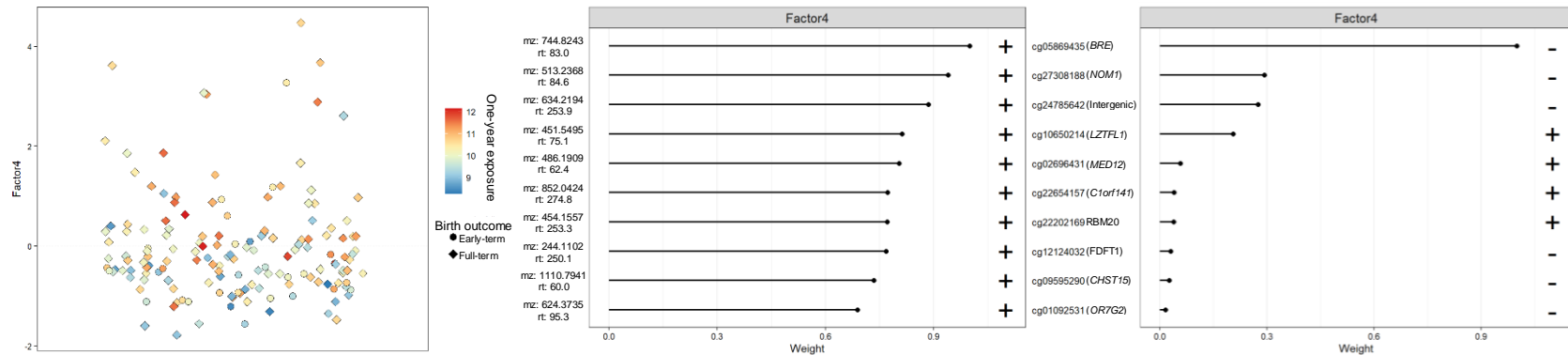
A. One-year exposure prior to conception with the mediating factor derived from 500 metabolic features of HILIC column and CpG sites



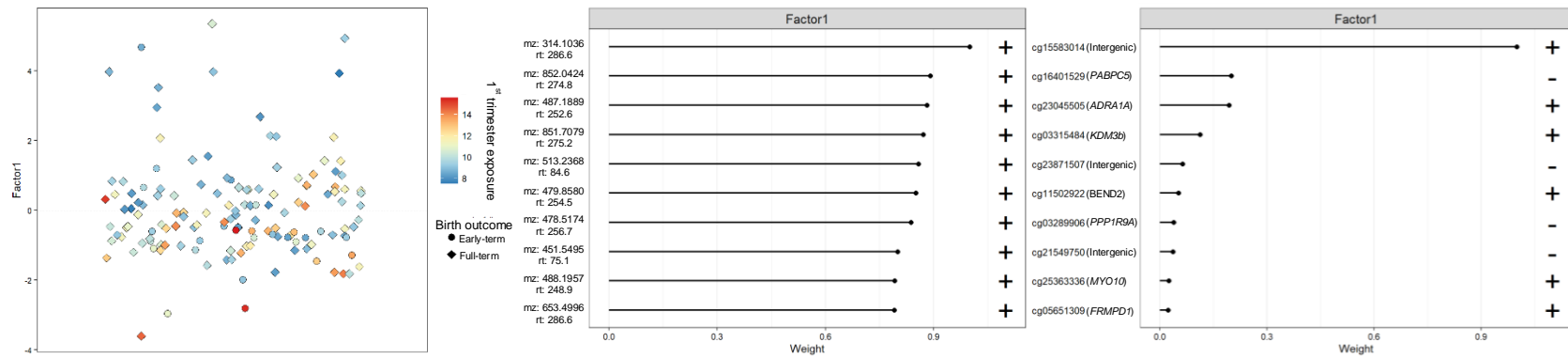
B. One-week exposure prior to blood draw with the mediating factor derived from 500 metabolic features of HILIC column and CpG sites



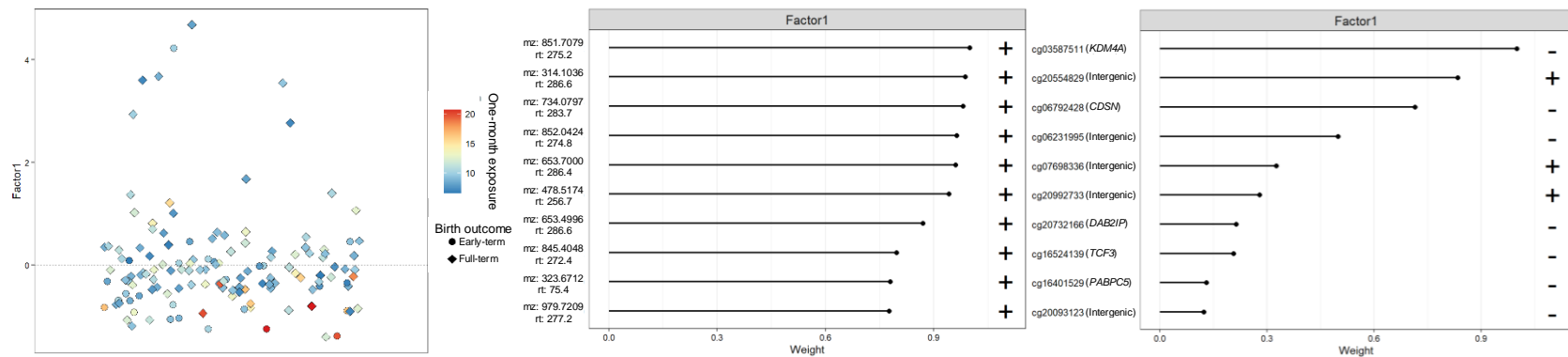
C. One-year exposure prior to conception with the mediating factor derived from 1000 metabolic features of HILIC column and CpG sites



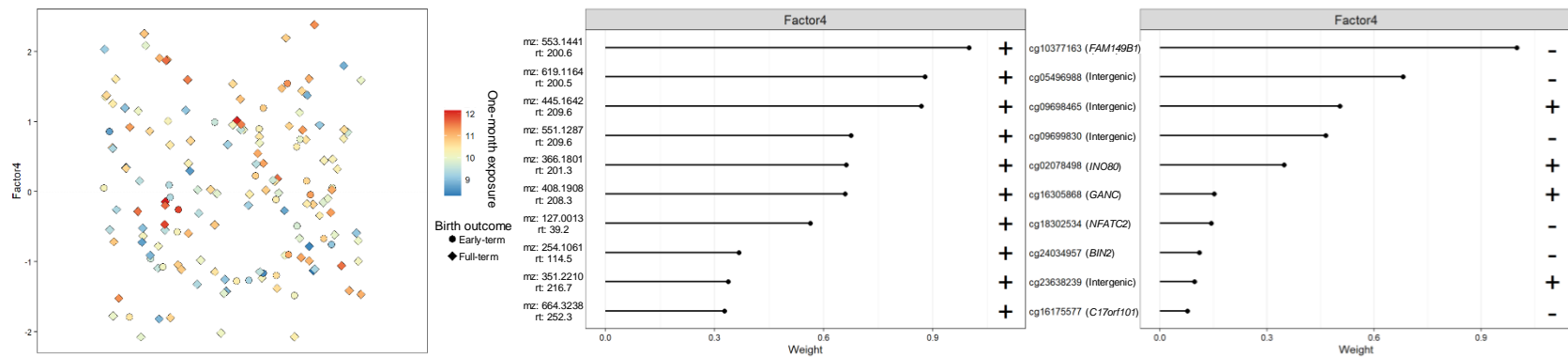
D. First-trimester exposure with the mediating factor derived from 1000 metabolic features of HILIC column and CpG sites



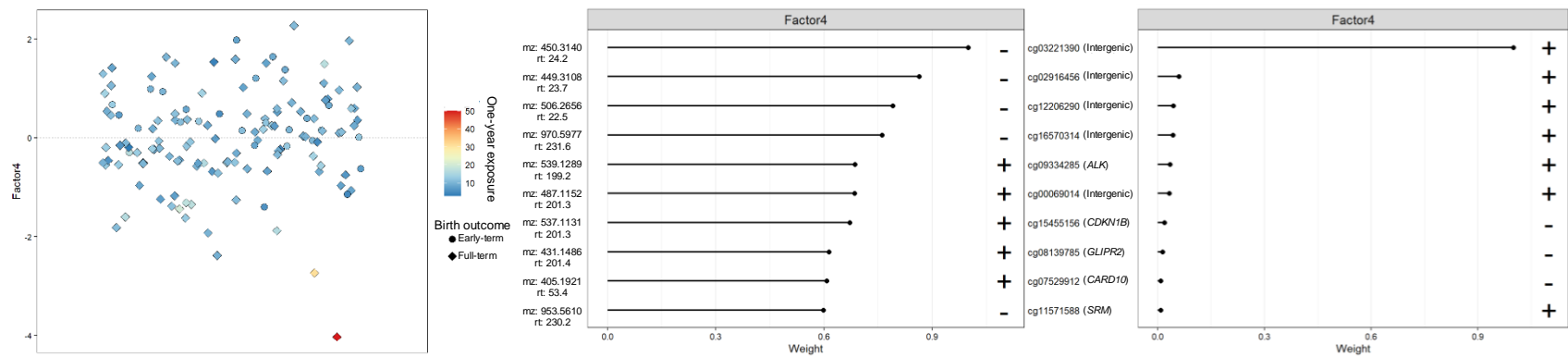
E. One-month exposure prior to blood draw with the mediating factor derived from 1000 metabolic features of HILIC column and CpG sites



F. One-year exposure prior to conception with the mediating factor derived from 500 metabolic features of C18 column and CpG sites



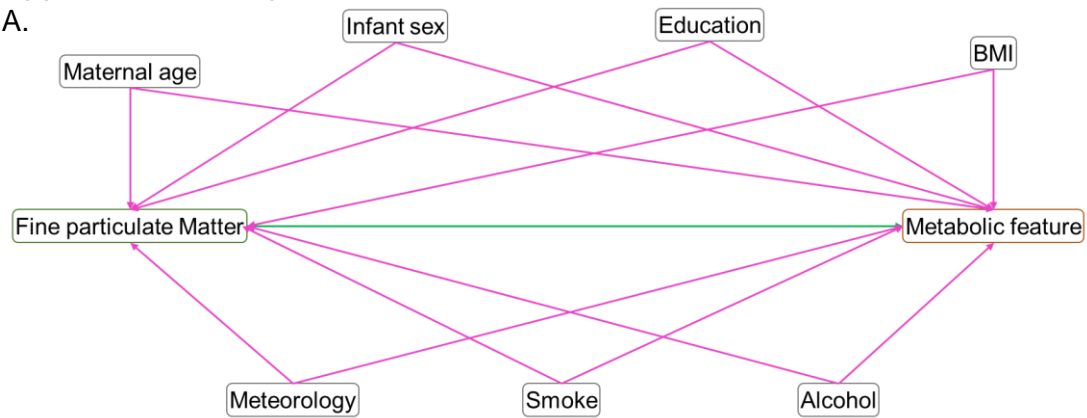
G. One-week exposure prior to blood draw with the mediating factor derived from 500 metabolic features of C18 column and CpG sites



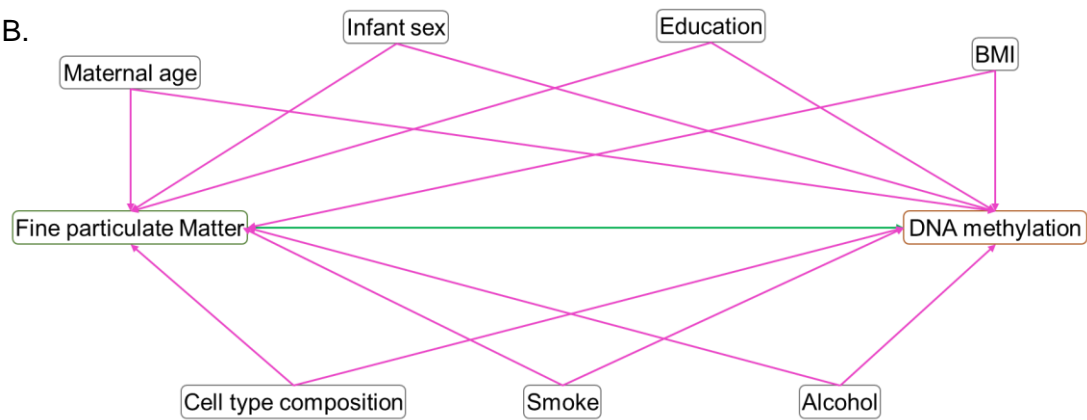
**Figure 2.** Visualization of mediating factors for the associations of fine particulate matter ( $PM_{2.5}$ ) exposures with early-term birth and the corresponding weights of the top 10 metabolic features and CpG sites for each factor. The plot on the left column illustrates the factor value, and each dot represents a participant colored by the  $PM_{2.5}$  exposure and shaped by the outcome. The middle column and right column show the weights of the metabolic features and CpG sites for the factor with the plus and minus signs indicating the direction of weights.

## Appendix for Chapter 5

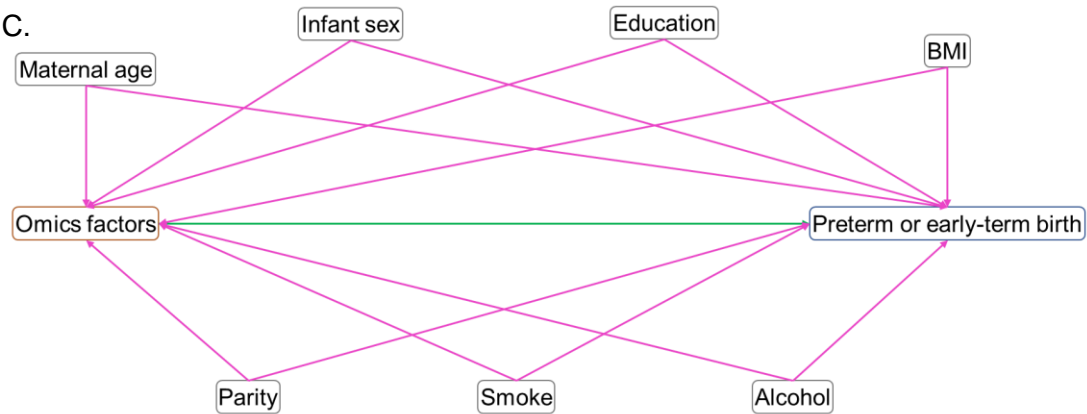
A.



B.



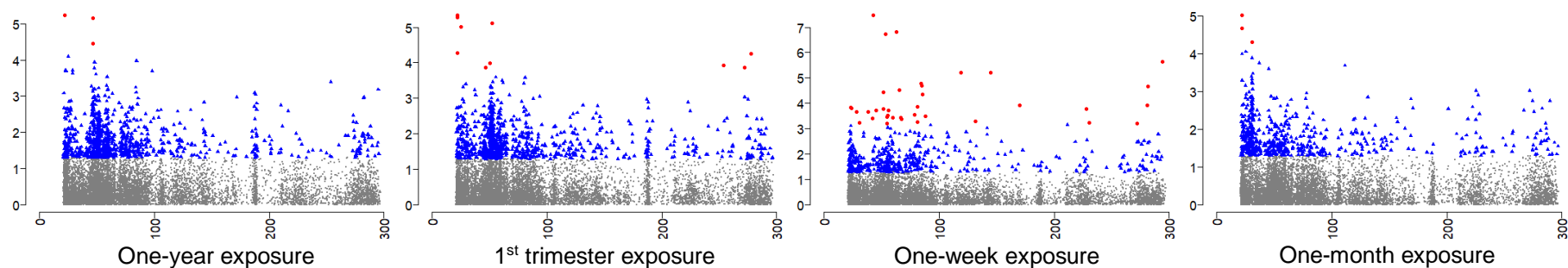
C.



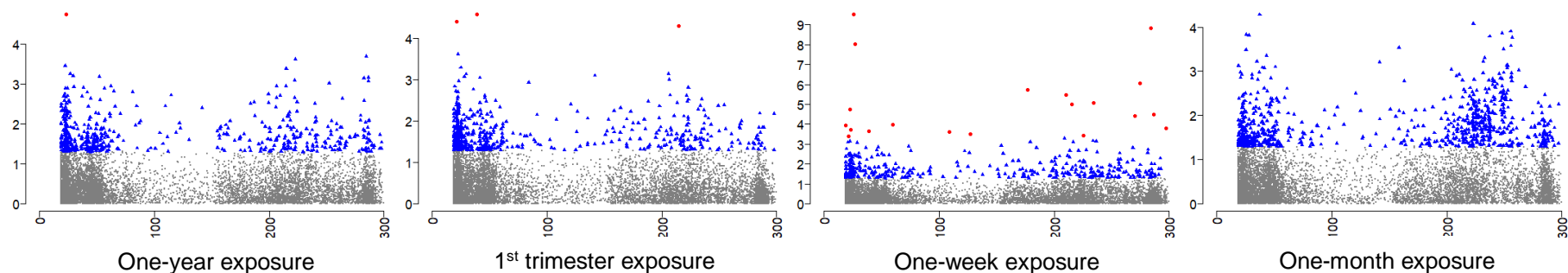
**Figure S1.** Directed acyclic graph of the confounding structure. BMI, body mass index; Smoke, tobacco and marijuana use in the month prior to pregnancy; Meteorology, conception season or averaged apparent temperature.



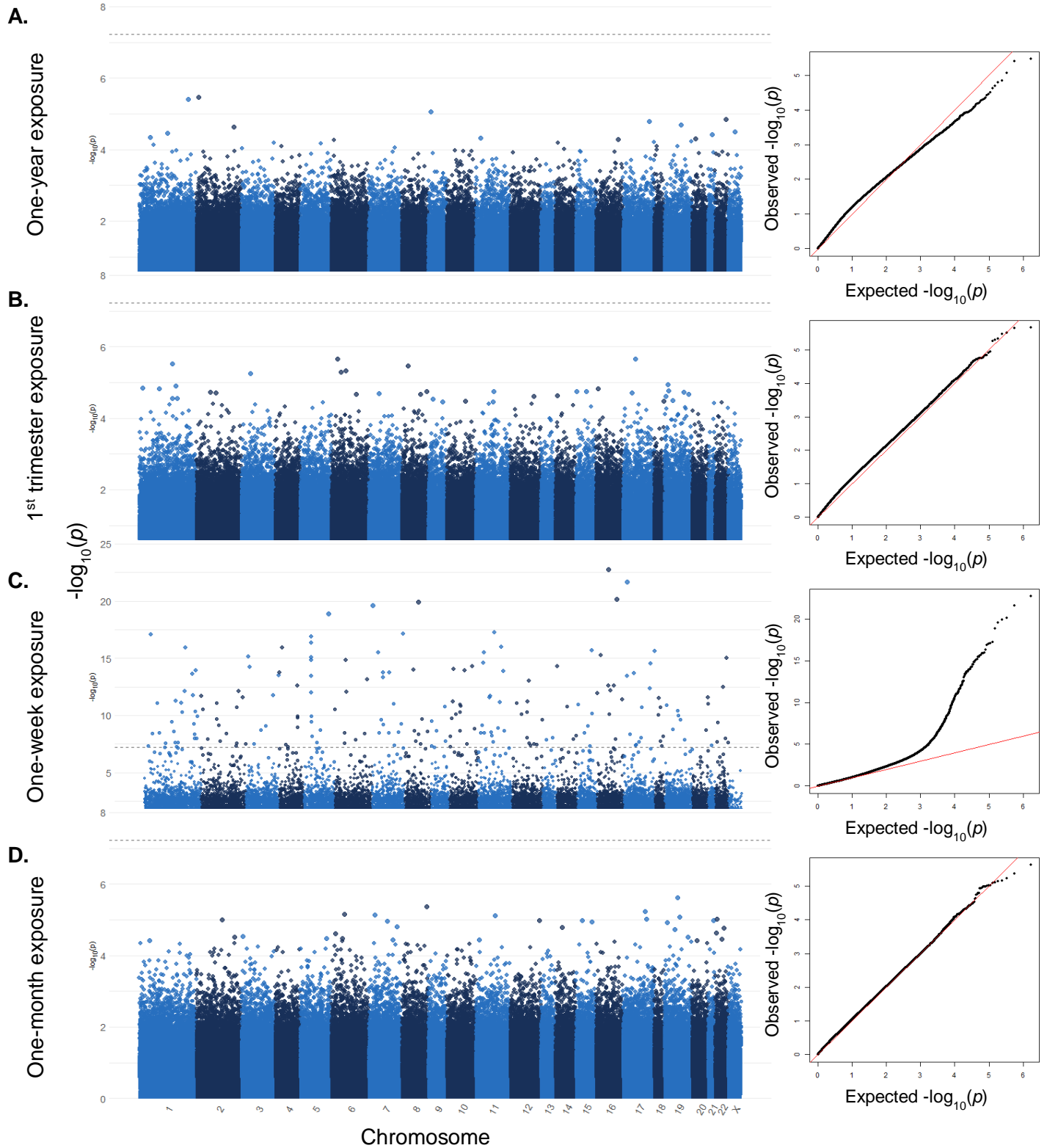
## A. HILIC column



## B. C18 column



**Figure S2.** Manhattan plots of metabolome-wide association analysis for  $PM_{2.5}$  exposures. A. Associations between  $PM_{2.5}$  exposures and changes in intensities of metabolic features in the HILIC column; B. Associations between  $PM_{2.5}$  exposures and changes in intensities of metabolic features in the C18 column. X-axis denotes the retention time (in seconds) of the metabolic features, and Y-axis denotes the negative  $\log_{10}$  of  $p$ -values. Red dots indicated significant associations at  $FDR_{B-H} < 0.2$ , and blue indicated associations at raw  $p$ -values  $< 0.05$ . Abbreviations: HILIC, hydrophilic interaction liquid chromatography; C18, hydrophobic reversed-phase chromatography;  $PM_{2.5}$ , fine particulate matter;  $FDR_{B-H}$ , Benjamini-Hochberg adjusted  $p$ -values.



**Figure S3.** Manhattan and QQ plots for the epigenome-wide association of PM<sub>2.5</sub> exposures (A. One-year prior to conception / B. 1<sup>st</sup> trimester / C. One-week prior to blood draw / D. One-month prior to blood draw average exposures) and DNAm in peripheral blood mononuclear cells. Bonferroni threshold: 0.05/ 839,487.

**Table S1.** Statistics of apparent temperature for the three short-term exposure windows among 264 pregnant people in the Atlanta African American Maternal-Child Cohort, 2014-2018.

Apparent temperature (Fahrenheit)	Q1	Q3	IQR	Median	Mean
1 <sup>st</sup> trimester	54.24	77.71	23.48	68.65	65.90
One-week prior to blood draw	54.88	78.50	23.62	66.60	65.08
One-month prior to blood draw	53.41	78.54	25.13	67.67	65.51

Abbreviations: Q1, 1<sup>st</sup> quartile; Q3, 3<sup>rd</sup> quartile; IQR, interquartile range.

**Table S2.** Number of metabolic features and differentially methylated CpG sites associated with PM<sub>2.5</sub> exposures under different significance cutoff among 264 pregnant people in the Atlanta African American Maternal-Child Cohort, 2014-2018.

Exposure	Outcome	Adjusted <i>p</i> -values <sup>a</sup>			Raw <i>p</i> -values		
		< 0.05	< 0.1	< 0.2	< 0.001	< 0.01	< 0.05
One-year	MWAS - HILIC column	2	2	3	33	277	957
	MWAS - C18 column	0	0	0	13	120	553
	EWAS	0	0	0	580	9782	63974
1 <sup>st</sup> trimester	MWAS - HILIC column	4	4	10	30	276	935
	MWAS - C18 column	0	0	3	11	110	534
	EWAS	0	0	0	1059	11490	60632
One-week	MWAS - HILIC column	12	14	37	43	140	472
	MWAS - C18 column	11	13	20	28	98	378
	EWAS	199	219	243	3034	14224	48329
One-month	MWAS - HILIC column	0	0	3	19	147	568
	MWAS - C18 column	0	0	0	26	209	713
	EWAS	0	0	0	816	8777	45764

Abbreviations: PM<sub>2.5</sub>, fine particulate matter; MWAS, metabolome-wide analysis study; EWAS, epigenome-wide analysis study;

HILIC, hydrophilic interaction liquid chromatography; C18, hydrophobic reversed-phase chromatography; FDR<sub>B-H</sub>, Benjamini-Hochberg adjusted *p*-values.

<sup>a</sup> For MWAS, raw *p*-values were adjusted by Benjamini-Hochberg approach; for EWAS, raw *p*-values were adjusted by Bonferroni approach.

## Chapter 6 Limitations, Implications and Future Direction

In this dissertation research, we used cutting-edge omics techniques to characterize biological changes in human body to illustrate the relationship between PM<sub>2.5</sub> exposures and health outcomes. Omics techniques, which include genomics, epigenomics, transcriptomics, proteomics, and metabolomics, have revolutionized the way we understand and study air pollution. With the help of these techniques, researchers can identify key molecules and pathways involved in the response to air pollution, which can provide new insights into the underlying mechanisms of pollution-related diseases.

In Chapter 2, I developed an air quality model that simulated weekly averaged PM<sub>2.5</sub> concentrations with a spatial resolution of 200m in Metropolitan Atlanta from 2012-2019, which has become the cornerstone of subsequent analyses. In Chapter 3, multiple CpG sites located in prefrontal cortex tissues were identified as potential mediators of the association between PM<sub>2.5</sub> exposure and neuropathology markers related to AD. Some of these CpG sites were situated in genes linked to neuroinflammation and neuroinflammation-mediated necroptosis in brain tissues, suggesting neuroinflammation as a potential underlying mechanism of PM<sub>2.5</sub> neurotoxicity. In Chapter 4, metabolomic signatures detected in early-pregnancy serum samples acted as potential mediators for the adverse effects of long- and short-term exposures to PM<sub>2.5</sub> on the risk of PTB and ETB. Specifically, biological pathways involved in folate metabolism and glycine and serine metabolism were found to play a pivotal role in the biological mechanism underlying PM<sub>2.5</sub> toxicity on early birth. In Chapter 5, by utilizing multi-omics integration techniques, latent factors were derived from metabolomic and epigenomic data, which were found to significantly mediate the association between PM<sub>2.5</sub> exposure and the risk of ETB. This provides valuable insights into the intricate interplay between metabolic features and DNA methylation, which appears to be critical in PM<sub>2.5</sub> toxicity.

These findings suggest that the application of omics techniques can provide a comprehensive understanding of the biological changes in the human body in response to PM<sub>2.5</sub> exposure and their role in disease development. By focusing on vulnerable populations such as older adults and pregnant women, this research sheds light on the molecular mechanisms of the pathogenesis of ETB and AD, which is crucial for risk assessment and the development of effective interventions for these populations.

The current findings are limited for a few reasons. First, the current project leveraged the spatial-temporal estimates of traffic-related and ambient PM<sub>2.5</sub> concentrations to derive exposure estimates from residential locations, as a surrogate for personal exposures, which is typically used for large populations or retrospective studies.<sup>229</sup> Due to the daily mobility and indoor PM<sub>2.5</sub> sources, the resulting measurement error might lead to underestimation or overestimation of the effect estimates, and the direction of bias can be hard to predict. Future studies using more accurate air pollution models or personal monitoring equipment are warranted to verify our findings. Second, PM<sub>2.5</sub> is composed of a complex mixture of chemical compounds, and the composition can affect its toxicity and potential health effects. The lack of information on the chemical composition limited our ability to evaluate the differences in the health impacts between traffic-related and ambient PM<sub>2.5</sub>. Third, both metabolomics and epigenomics are a dynamic system, but only a snapshot of metabolomic or epigenomic profiles was included in the Chapters 3-5. Failing to account for their dynamic nature might lead to incorrect or incomplete interpretation of results and inability to capture the full complexity of biological processes. Thus, it is important to consider the temporal aspect of biological processes, and integrating the same type of omics data at multiple time points should be an integral part of multi-omics integration as well. Forth, the multi-omics tools are often developed and optimized for specific types of omics data, outcomes or study designs. Although I optimized the workflow in Chapter 5 to fit *MOFA2* and *mixOmics* in the current context, the findings of multi-omics integration must be evaluated with caution when generalizing them to different

contexts or populations. Further research is needed to develop more flexible and adaptable multi-omics tools.

In the future, several areas warranted further investigation in air pollution research using omics techniques: 1) Health effects of mixtures: Air pollution is not composed of a single pollutant, but rather of complex mixtures of pollutants. Omics techniques can be used to study the health effects of these mixtures, and to identify the key pollutants and pathways that are responsible for the observed effects. New statistical methods are needed to deal with the more complex relationships between mixtures and high-dimensional omics data. 2) Mediating role of omics data: to our best knowledge, although a few high-dimensional analysis tools have been published, they employed different frameworks which might lead to inconsistent results. Moreover, there is no such a widely-accepted workflow or method that works for all omics types. The development of high-dimensional analysis for specific omics data and study design is still necessary. 3) Multi-omics integration in environmental epidemiology. As most environmental epidemiological studies are observational, confounding is always a crucial consideration. A multi-omics integration workflow that fits environmental epidemiological studies must take into account the confounding factors.

Overall, omics techniques will continue to play a critical role in air pollution research in the coming years, providing new insights into the health effects of air pollution and new approaches for preventing and treating pollution-related diseases.

## References

1. Sarnat JA, Russell A, Liang D, et al. Developing Multipollutant Exposure Indicators of Traffic Pollution: The Dorm Room Inhalation to Vehicle Emissions (DRIVE) Study. *Res Rep Health Eff Inst.* Apr 2018;(196):3-75.
2. Liang CS, Duan FK, He KB, Ma YL. Review on recent progress in observations, source identifications and countermeasures of PM<sub>2.5</sub>. *Environ Int.* Jan 2016;86:150-70.  
doi:10.1016/j.envint.2015.10.016
3. Fox M, Koehler K, Johnson N. Established and emerging effects of traffic-related air pollution. *Traffic-Related Air Pollution*. Elsevier; 2020:207-228.
4. Zhu X, Ji X, Shou Y, Huang Y, Hu Y, Wang H. Recent advances in understanding the mechanisms of PM<sub>2.5</sub>-mediated neurodegenerative diseases. *Toxicol Lett.* Sep 1 2020;329:31-37. doi:10.1016/j.toxlet.2020.04.017
5. Miller MR, Raftis JB. Evidence from toxicological and mechanistic studies. *Traffic-Related Air Pollution*. Elsevier; 2020:229-279.
6. Wang Y, Zhang M, Li Z, et al. Fine particulate matter induces mitochondrial dysfunction and oxidative stress in human SH-SY5Y cells. *Chemosphere.* Mar 2019;218:577-588.  
doi:10.1016/j.chemosphere.2018.11.149
7. Wang BR, Shi JQ, Ge NN, et al. PM<sub>2.5</sub> exposure aggravates oligomeric amyloid beta-induced neuronal injury and promotes NLRP3 inflammasome activation in an in vitro model of Alzheimer's disease. *J Neuroinflammation.* May 2 2018;15(1):132. doi:10.1186/s12974-018-1178-5
8. Bhatt DP, Puig KL, Gorr MW, Wold LE, Combs CK. A pilot study to assess effects of long-term inhalation of airborne particulate matter on early Alzheimer-like changes in the mouse brain. *PLoS One.* 2015;10(5):e0127102. doi:10.1371/journal.pone.0127102



9. Tyler CR, Zychowski KE, Sanchez BN, et al. Surface area-dependence of gas-particle interactions influences pulmonary and neuroinflammatory outcomes. *Part Fibre Toxicol.* Dec 1 2016;13(1):64. doi:10.1186/s12989-016-0177-x
10. Li Z, Tang Y, Song X, Lazar L, Li Z, Zhao J. Impact of ambient PM<sub>2.5</sub> on adverse birth outcome and potential molecular mechanism. *Ecotoxicol Environ Saf.* Mar 2019;169:248-254. doi:10.1016/j.ecoenv.2018.10.109
11. Nagiah S, Phulukdaree A, Naidoo D, et al. Oxidative stress and air pollution exposure during pregnancy: A molecular assessment. *Hum Exp Toxicol.* Aug 2015;34(8):838-47. doi:10.1177/0960327114559992
12. de Melo JO, Soto SF, Katayama IA, et al. Inhalation of fine particulate matter during pregnancy increased IL-4 cytokine levels in the fetal portion of the placenta. *Toxicol Lett.* Jan 22 2015;232(2):475-80. doi:10.1016/j.toxlet.2014.12.001
13. Zhang M, Liu W, Zhou Y, Li Y, Qin Y, Xu Y. Neurodevelopmental toxicity induced by maternal PM<sub>2.5</sub> exposure and protective effects of quercetin and Vitamin C. *Chemosphere.* Dec 2018;213:182-196. doi:10.1016/j.chemosphere.2018.09.009
14. Nachman RM, Mao G, Zhang X, et al. Intrauterine Inflammation and Maternal Exposure to Ambient PM<sub>2.5</sub> during Preconception and Specific Periods of Pregnancy: The Boston Birth Cohort. *Environ Health Perspect.* Oct 2016;124(10):1608-1615. doi:10.1289/EHP243
15. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol.* Dec 1986;51(6):1173-82. doi:10.1037//0022-3514.51.6.1173
16. Pinu FR, Beale DJ, Paten AM, et al. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites.* Apr 18 2019;9(4)doi:10.3390/metabo9040076

17. Liang D, Li Z, Vlaanderen J, Tang Z, Vermeulen R, Sarnat J. Systematic Review on Untargeted Metabolomics Application in Air Pollution Health Research: Current Progress, Analytical Challenges, and Future Direction. 2020:
18. Sun YV, Hu YJ. Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. *Adv Genet.* 2016;93:147-90.  
doi:10.1016/bs.adgen.2015.11.004
19. Anderson OS, Sant KE, Dolinoy DC. Nutrition and epigenetics: an interplay of dietary methyl donors, one-carbon metabolism and DNA methylation. *J Nutr Biochem.* Aug 2012;23(8):853-9.  
doi:10.1016/j.jnutbio.2012.03.003
20. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. *Hum Mol Genet.* Oct 1 2016;25(R2):R149-R156. doi:10.1093/hmg/ddw197
21. Aung MT, Song Y, Ferguson KK, et al. Application of an analytical framework for multivariate mediation analysis of environmental data. *Nat Commun.* Nov 6 2020;11(1):5624.  
doi:10.1038/s41467-020-19335-2
22. Zhao Y, Lindquist MA, Caffo BS. Sparse principal component based high-dimensional mediation analysis. *Computational statistics & data analysis.* 2020;142:106835.
23. Zhang H, Zheng Y, Zhang Z, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics.* 2016;32(20):3150-3154.
24. Yang T, Niu J, Chen H, Wei P. Estimation of Mediation Effect for High-dimensional Omics Mediators with Application to the Framingham Heart Study. *bioRxiv.* 2019:774877.
25. Huang YT, Pan WC. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics.* Jun 2016;72(2):402-13.  
doi:10.1111/biom.12421
26. Song Y, Zhou X, Zhang M, et al. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics.* 2020;76(3):700-710.

27. Inoue K, Yan Q, Arah OA, et al. Air Pollution and Adverse Pregnancy and Birth Outcomes: Mediation Analysis Using Metabolomic Profiles. *Curr Environ Health Rep*. Sep 2020;7(3):231-242. doi:10.1007/s40572-020-00284-3
28. Lin VW, Baccarelli AA, Burris HH. Epigenetics—a potential mediator between air pollution and preterm birth. *Environmental epigenetics*. 2016;2(1)
29. Chadeau-Hyam M, Athersuch TJ, Keun HC, et al. Meeting-in-the-middle using metabolic profiling—a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*. 2011;16(1):83-88.
30. Bachmann J. Will the circle be unbroken: a history of the U.S. National Ambient Air Quality Standards. *J Air Waste Manag Assoc*. Jun 2007;57(6):652-97. doi:10.3155/1047-3289.57.6.652
31. Askariyeh MH, Khreis H, Vallamsundar S. Air pollution monitoring and modeling. *Traffic-Related Air Pollution*. Elsevier; 2020:111-135.
32. Yatkin S, Gerboles M, Belis CA, et al. Representativeness of an air quality monitoring station for PM(2.5) and source apportionment over a small urban domain. *Atmos Pollut Res*. Feb 2020;11(2):225-233. doi:10.1016/j.apr.2019.10.004
33. Wilson JG, Kingham S, Pearce J, Sturman AP. A review of intraurban variations in particulate air pollution: Implications for epidemiological research. *Atmos Environ*. Nov 2005;39(34):6444-6462. doi:10.1016/j.atmosenv.2005.07.030
34. D'Onofrio D, Kim B, Kim Y, Kim K. Atlanta Roadside Emissions Exposure Study- Methodology & Project Overview. Atlanta Regional Commission; 2016.
35. Zhai X, Russell AG, Sampath P, et al. Calibrating R-LINE model results with observational data to develop annual mobile source air pollutant fields at fine spatial resolution: Application in Atlanta. *Atmos Environ*. 2016;147:446-457.
36. Bates JT, Pennington AF, Zhai XX, et al. Application and evaluation of two model fusion approaches to obtain ambient air pollutant concentrations at a fine spatial resolution (250m) in Atlanta. *Environ Modell Softw*. Nov 2018;109:182-190. doi:10.1016/j.envsoft.2018.06.008

37. Cimorelli AJ, Perry SG, Venkatram A, et al. AERMOD: A dispersion model for industrial source applications. Part I: General model formulation and boundary layer characterization. *J Appl Meteorol*. May 2005;44(5):682-693. doi:Doi 10.1175/Jam2227.1
38. Rousseau G, Clymer T. Travel demand modeling and conformity determination - Atlanta Regional Commission case study. *Transport Res Rec*. 2002;(1817):172-176.
39. U.S. Environmental Protection Agency. User Guide for MOVES2014. 1/29/2023, 2023. Updated 1/4/2023. Accessed 1/29/2023, 2023. <https://www.epa.gov/moves/moves-versions-limited-current-use>
40. Georgia Department of Transportation. Road & Traffic Data. 1/29/2023, Accessed 8/1/2022, 2022. <https://www.dot.ga.gov/GDOT/Pages/RoadTrafficData.aspx>
41. Homer C, Dewitz J, Jin SM, et al. Conterminous United States land cover change patterns 2001-2016 from the 2016 National Land Cover Database. *Isprs J Photogramm*. Apr 2020;162:184-199. doi:10.1016/j.isprsjprs.2020.02.019
42. Multi-Resolution Land Characteristics Consortium. National Land Cover Database Class Legend and Description. 1/29/2023, 2023. Accessed 5/08/2022, 2021.
43. van Donkelaar A, Hammer MS, Bindle L, et al. Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environ Sci Technol*. Nov 16 2021;55(22):15287-15300. doi:10.1021/acs.est.1c05309
44. RColorBrewer S, Liaw MA. Package 'randomforest'. *University of California, Berkeley: Berkeley, CA, USA*. 2018;
45. Binkowski FS, Roselle SJ. Models-3 community multiscale air quality (CMAQ) model aerosol component - 1. Model description. *J Geophys Res-Atmos*. Mar 26 2003;108(D6)doi:Artn 4183 10.1029/2001jd001409
46. Berrocal VJ, Gelfand AE, Holland DM. A Spatio-Temporal Downscaler for Output From Numerical Models. *J Agr Biol Envir St*. Jun 2010;15(2):176-197. doi:10.1007/s13253-009-0004-

47. Holland D. Fused Air Quality Predictions Using Downscaling. *Research Triangle Park, NC US Environmental Protection Agency* [https://www.epa.gov/sites/production/files/2016-07/documents/data\\_fusion\\_meta\\_file\\_july\\_2016.pdf](https://www.epa.gov/sites/production/files/2016-07/documents/data_fusion_meta_file_july_2016.pdf) [accessed 17 January 2014]. 2012;
48. Akima H. A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures. *J Acm.* 1970;17(4):589-&. doi:Doi 10.1145/321607.321609
49. Goldman GT, Mulholland JA, Russell AG, et al. Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies. *Environ Health.* Jun 22 2011;10:61. doi:10.1186/1476-069X-10-61
50. Sheppard L, Burnett RT, Szpiro AA, et al. Confounding and exposure measurement error in air pollution epidemiology. *Air Qual Atmos Health.* Jun 2012;5(2):203-216. doi:10.1007/s11869-011-0140-9
51. Snyder MG, Venkatram A, Heist DK, Perry SG, Petersen WB, Isakov V. RLINE: A line source dispersion model for near-surface releases. *Atmospheric environment.* 2013;77:748-756.
52. Khreis H, Nieuwenhuijsen MJ, Zietsman J, Ramani T. Traffic-related air pollution: Emissions, human exposures, and health: An introduction. *Traffic-related air pollution.* Elsevier; 2020:1-21.
53. Harrison RM, Jones AM, Lawrence RG. Major component composition of PM<sub>10</sub> and PM<sub>2.5</sub> from roadside and urban background sites. *Atmospheric Environment.* 2004;38(27):4531-4538.
54. Tessum MW, Anenberg SC, Chafe ZA, et al. Sources of ambient PM<sub>2.5</sub> exposure in 96 global cities. *Atmos Environ (1994).* Oct 1 2022;286:119234. doi:10.1016/j.atmosenv.2022.119234
55. Jeong CH, Wang JM, Hilker N, et al. Temporal and spatial variability of traffic-related PM<sub>2.5</sub> sources: Comparison of exhaust and non-exhaust emissions. *Atmos Environ.* Feb 1 2019;198:55-69. doi:10.1016/j.atmosenv.2018.10.038
56. Park M, Joo HS, Lee K, et al. Differential toxicities of fine particulate matters from various sources. *Sci Rep.* Nov 19 2018;8(1):17007. doi:10.1038/s41598-018-35398-0

57. Fu PF, Guo XB, Cheung FMH, Yung KKL. The association between PM2.5 exposure and neurological disorders: A systematic review and meta-analysis. *Sci Total Environ*. Mar 10 2019;655:1240-1248. doi:10.1016/j.scitotenv.2018.11.218
58. Shou YK, Huang YL, Zhu XZ, Liu CQ, Hu Y, Wang HH. A review of the possible associations between ambient PM2.5 exposures and the development of Alzheimer's disease. *Ecotox Environ Safe*. Jun 15 2019;174:344-352. doi:10.1016/j.ecoenv.2019.02.086
59. Alzheimer's Association. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement*. Apr 2022;18(4):700-789. doi:10.1002/alz.12638
60. Brookmeyer R, Abdalla N, Kawas CH, Corrada MM. Forecasting the prevalence of preclinical and clinical Alzheimer's disease in the United States. *Alzheimers Dement*. Feb 2018;14(2):121-129. doi:10.1016/j.jalz.2017.10.009
61. Deb A, Thornton JD, Sambamoorthi U, Innes K. Direct and indirect cost of managing alzheimer's disease and related dementias in the United States. *Expert Rev Pharmacoecon Outcomes Res*. Apr 2017;17(2):189-202. doi:10.1080/14737167.2017.1313118
62. Killin LO, Starr JM, Shiue IJ, Russ TC. Environmental risk factors for dementia: a systematic review. *BMC Geriatr*. Oct 12 2016;16(1):175. doi:10.1186/s12877-016-0342-y
63. Kang YJ, Tan HY, Lee CY, Cho H. An Air Particulate Pollutant Induces Neuroinflammation and Neurodegeneration in Human Brain Models. *Adv Sci (Weinh)*. Nov 2021;8(21):e2101251. doi:10.1002/adv.202101251
64. Kilian J, Kitazawa M. The emerging risk of exposure to air pollution on cognitive decline and Alzheimer's disease - Evidence from epidemiological and animal studies. *Biomed J*. Jun 2018;41(3):141-162. doi:10.1016/j.bj.2018.06.001
65. Yokoyama AS, Rutledge JC, Medici V. DNA methylation alterations in Alzheimer's disease. *Environ Epigenet*. May 2017;3(2):dvx008. doi:10.1093/eep/dvx008

66. Iwata A, Nagata K, Hatsuta H, et al. Altered CpG methylation in sporadic Alzheimer's disease is associated with APP and MAPT dysregulation. *Hum Mol Genet.* Feb 1 2014;23(3):648-56. doi:10.1093/hmg/ddt451
67. Wang SC, Oelze B, Schumacher A. Age-specific epigenetic drift in late-onset Alzheimer's disease. *PLoS One.* Jul 16 2008;3(7):e2698. doi:10.1371/journal.pone.0002698
68. Smith RG, Hannon E, De Jager PL, et al. Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. *Alzheimers & Dementia.* Dec 2018;14(12):1580-1588. doi:10.1016/j.jalz.2018.01.017
69. Nicolia V, Cavallaro RA, Lopez-Gonzalez I, et al. DNA Methylation Profiles of Selected Pro-Inflammatory Cytokines in Alzheimer Disease. *J Neuropathol Exp Neurol.* Jan 1 2017;76(1):27-31. doi:10.1093/jnen/nlw099
70. Huls A, Robins C, Conneely KN, et al. Brain DNA Methylation Patterns in CLDN5 Associated With Cognitive Decline. *Biol Psychiatry.* Feb 15 2022;91(4):389-398. doi:10.1016/j.biopsych.2021.01.015
71. Wu Y, Qie R, Cheng M, et al. Air pollution and DNA methylation in adults: A systematic review and meta-analysis of observational studies. *Environ Pollut.* Sep 1 2021;284:117152. doi:10.1016/j.envpol.2021.117152
72. Newcombe EA, Camats-Perna J, Silva ML, Valmas N, Huat TJ, Medeiros R. Inflammation: the link between comorbidities, genetics, and Alzheimer's disease. *J Neuroinflamm.* Sep 24 2018;15doi:ARTN 27610.1186/s12974-018-1313-3
73. Fang J, Gao Y, Zhang M, et al. Personal PM(2.5) Elemental Components, Decline of Lung Function, and the Role of DNA Methylation on Inflammation-Related Genes in Older Adults: Results and Implications of the BAPE Study. *Environ Sci Technol.* Nov 15 2022;56(22):15990-16000. doi:10.1021/acs.est.2c04972

74. Tachibana K, Takayanagi K, Akimoto A, et al. Prenatal diesel exhaust exposure disrupts the DNA methylation profile in the brain of mouse offspring. *J Toxicol Sci.* Feb 2015;40(1):1-11. doi:10.2131/jts.40.1
75. Wei H, Liang F, Meng G, et al. Redox/methylation mediated abnormal DNA methylation as regulators of ambient fine particulate matter-induced neurodevelopment related impairment in human neuronal cells. *Sci Rep.* Sep 14 2016;6:33402. doi:10.1038/srep33402
76. Besser LM, Kukull WA, Teylan MA, et al. The Revised National Alzheimer's Coordinating Center's Neuropathology Form-Available Data and New Analyses. *J Neuropathol Exp Neurol.* Aug 1 2018;77(8):717-726. doi:10.1093/jnen/nly049
77. Mirra SS, Heyman A, McKeel D, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology.* Apr 1991;41(4):479-86. doi:10.1212/wnl.41.4.479
78. Hyman BT, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers Dement.* Jan 2012;8(1):1-13. doi:10.1016/j.jalz.2011.10.007
79. Thal DR, Rub U, Orantes M, Braak H. Phases of A beta-deposition in the human brain and its relevance for the development of AD. *Neurology.* Jun 25 2002;58(12):1791-800. doi:10.1212/wnl.58.12.1791
80. Manual I. Georgia Department of Transportation. *Office of Materials and Testing, Atlanta, GA.* 2019;
81. Homer CH, Fry JA, Barnes CA. The national land cover database. *US geological survey fact sheet.* 2012;3020(4):1-4.
82. van Donkelaar A, Hammer MS, Bindle L, et al. Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environ Sci Technol.* Nov 16 2021;55(22):15287-15300. doi:10.1021/acs.est.1c05309



83. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. Jan 1 2012;28(1):112-8. doi:10.1093/bioinformatics/btr597
84. Konwar C, Price EM, Wang LQ, Wilson SL, Terry J, Robinson WP. DNA methylation profiling of acute chorioamnionitis-associated placentas and fetal membranes: insights into epigenetic variation in spontaneous preterm births. *Epigenetics Chromatin*. Oct 29 2018;11(1):63. doi:10.1186/s13072-018-0234-9
85. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. Jan 2007;8(1):118-27. doi:10.1093/biostatistics/kxj037
86. Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*. Mar 2013;8(3):290-302. doi:10.4161/epi.23924
87. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. May 15 2014;30(10):1363-9. doi:10.1093/bioinformatics/btu049
88. Kind AJH, Buckingham WR. Making Neighborhood-Disadvantage Metrics Accessible - The Neighborhood Atlas. *N Engl J Med*. Jun 28 2018;378(26):2456-2458. doi:10.1056/NEJMp1802313
89. Venables WN, Ripley BD. *Modern applied statistics with S-PLUS*. Springer Science & Business Media; 2013.
90. Leek JT, Johnson WE, Parker HS, et al. sva: Surrogate variable analysis. *R package version*. 2019;3(0):882-883.
91. van Iterson M, van Zwet EW, Consortium B, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol*. Jan 27 2017;18(1):19. doi:10.1186/s13059-016-1131-9
92. Armstrong RA. When to use the Bonferroni correction. *Ophthalm Physiol Opt*. Sep 2014;34(5):502-508. doi:10.1111/opo.12131

93. Chadeau-Hyam M, Athersuch TJ, Keun HC, et al. Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*. Feb 2011;16(1):83-8. doi:10.3109/1354750X.2010.533285
94. Zhang H, Zheng Y, Zhang Z, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*. Oct 15 2016;32(20):3150-3154. doi:10.1093/bioinformatics/btw351
95. Liu ZH, Shen JC, Barfield R, Schwartz J, Baccarelli AA, Lin XH. Large-Scale Hypothesis Testing for Causal Mediation Effects with Applications in Genome-wide Epigenetic Studies. *J Am Stat Assoc*. Jan 2 2022;117(537):67-81. doi:10.1080/01621459.2021.1914634
96. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. *J Stat Softw*. Aug 2014;59(5)
97. Kilanowski A, Merid SK, Abrishamcar S, et al. DNA methylation and aeroallergen sensitization: The chicken or the egg? *Clin Epigenetics*. Sep 16 2022;14(1):114. doi:10.1186/s13148-022-01332-5
98. Abrishamcar S, Chen JY, Feil D, et al. DNA methylation as a potential mediator of the association between prenatal tobacco and alcohol exposure and child neurodevelopment in a South African birth cohort. *Transl Psychiat*. Sep 30 2022;12(1)doi:ARTN 418 10.1038/s41398-022-02195-3
99. Maksimovic J, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biology*. Jun 8 2021;22(1)doi:ARTN 173 10.1186/s13059-021-02388-x
100. Hansen K. IlluminaHumanMethylationEPICanno. ilm10b2. hg19: annotation for Illumina's EPIC methylation arrays. R package version 0.6. 0. 2016.
101. Battram T, Yousefi P, Crawford G, et al. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res*. 2022;7:41. doi:10.12688/wellcomeopenres.17598.2

102. Min JL, Hemani G, Hannon E, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet.* Sep 2021;53(9):1311-1321.  
doi:10.1038/s41588-021-00923-x
103. Eisenberg DT, Kuzawa CW, Hayes MG. Worldwide allele frequencies of the human apolipoprotein E gene: climate, local adaptations, and evolutionary history. *Am J Phys Anthropol.* Sep 2010;143(1):100-11. doi:10.1002/ajpa.21298
104. Sordillo JE, Cardenas A, Qi C, et al. Residential PM(2.5) exposure and the nasal methylome in children. *Environ Int.* Aug 2021;153:106505. doi:10.1016/j.envint.2021.106505
105. Liu ML, Zang F, Zhang SJ. RBCK1 contributes to chemoresistance and stemness in colorectal cancer (CRC). *Biomed Pharmacother.* Oct 2019;118:109250.  
doi:10.1016/j.biopha.2019.109250
106. Niu Z, Fan J, Chen F, et al. RBCK1 regulates the progression of ER-positive breast cancer through the HIF1alpha signaling. *Cell Death Dis.* Dec 6 2022;13(12):1023.  
doi:10.1038/s41419-022-05473-6
107. Yu S, Dai J, Ma M, et al. RBCK1 promotes p53 degradation via ubiquitination in renal cell carcinoma. *Cell Death Dis.* Mar 15 2019;10(4):254. doi:10.1038/s41419-019-1488-2
108. Jazvinscak Jembrek M, Slade N, Hof PR, Simic G. The interactions of p53 with tau and Ass as potential therapeutic targets for Alzheimer's disease. *Prog Neurobiol.* Sep 2018;168:104-127. doi:10.1016/j.pneurobio.2018.05.001
109. Taminiau A, Draime A, Tys J, et al. HOXA1 binds RBCK1/HOIL-1 and TRAF2 and modulates the TNF/NF-kappaB pathway in a transcription-independent manner. *Nucleic Acids Res.* Sep 6 2016;44(15):7331-49. doi:10.1093/nar/gkw606
110. Elton L, Carpentier I, Verhelst K, Staal J, Beyaert R. The multifaceted role of the E3 ubiquitin ligase HOIL-1: beyond linear ubiquitination. *Immunol Rev.* Jul 2015;266(1):208-21.  
doi:10.1111/imr.12307

111. Srinivasan M, Lahiri DK. Significance of NF-kappaB as a pivotal therapeutic target in the neurodegenerative pathologies of Alzheimer's disease and multiple sclerosis. *Expert Opin Ther Targets*. Apr 2015;19(4):471-87. doi:10.1517/14728222.2014.989834
112. Snow WM, Albeni BC. Neuronal Gene Targets of NF-kappaB and Their Dysregulation in Alzheimer's Disease. *Front Mol Neurosci*. 2016;9:118. doi:10.3389/fnmol.2016.00118
113. Caccamo A, Branca C, Piras IS, et al. Necroptosis activation in Alzheimer's disease. *Nat Neurosci*. Sep 2017;20(9):1236-1246. doi:10.1038/nn.4608
114. Jayaraman A, Htike TT, James R, Picon C, Reynolds R. TNF-mediated neuroinflammation is linked to neuronal necroptosis in Alzheimer's disease hippocampus. *Acta Neuropathol Commun*. Sep 28 2021;9(1):159. doi:10.1186/s40478-021-01264-w
115. Shigemizu D, Asanomi Y, Akiyama S, Mitsumori R, Niida S, Ozaki K. Whole-genome sequencing reveals novel ethnicity-specific rare variants associated with Alzheimer's disease. *Mol Psychiatry*. May 2022;27(5):2554-2562. doi:10.1038/s41380-022-01483-0
116. Wang B, Bao S, Zhang Z, et al. A rare variant in MLKL confers susceptibility to APOE varepsilon4-negative Alzheimer's disease in Hong Kong Chinese population. *Neurobiol Aging*. Aug 2018;68:160 e1-160 e7. doi:10.1016/j.neurobiolaging.2018.03.006
117. Janelidze S, Zetterberg H, Mattsson N, et al. CSF Abeta42/Abeta40 and Abeta42/Abeta38 ratios: better diagnostic markers of Alzheimer disease. *Ann Clin Transl Neurol*. Mar 2016;3(3):154-65. doi:10.1002/acn3.274
118. Chen S, Liu H, Wang S, et al. The Neuroprotection of Verbascoside in Alzheimer's Disease Mediated through Mitigation of Neuroinflammation via Blocking NF-kappaB-p65 Signaling. *Nutrients*. Mar 29 2022;14(7)doi:10.3390/nu14071417
119. Lee JH, Cheng R, Vardarajan B, et al. Genetic Modifiers of Age at Onset in Carriers of the G206A Mutation in PSEN1 With Familial Alzheimer Disease Among Caribbean Hispanics. *JAMA Neurol*. Sep 2015;72(9):1043-51. doi:10.1001/jamaneurol.2015.1424

120. Chao MW, Yang CH, Lin PT, et al. Exposure to PM(2.5) causes genetic changes in fetal rat cerebral cortex and hippocampus. *Environ Toxicol.* Apr 2017;32(4):1412-1425.  
doi:10.1002/tox.22335
121. Vdovenko D, Bachmann M, Wijnen WJ, Hottiger MO, Eriksson U, Valaperti A. The adaptor protein c-Cbl-associated protein (CAP) limits pro-inflammatory cytokine expression by inhibiting the NF-kappaB pathway. *Int Immunopharmacol.* Oct 2020;87:106822.  
doi:10.1016/j.intimp.2020.106822
122. Heckman PR, Wouters C, Prickaerts J. Phosphodiesterase inhibitors as a target for cognition enhancement in aging and Alzheimer's disease: a translational overview. *Curr Pharm Des.* 2015;21(3):317-31. doi:10.2174/1381612820666140826114601
123. Puthiyedth N, Riveros C, Berretta R, Moscato P. Identification of Differentially Expressed Genes through Integrated Study of Alzheimer's Disease Affected Brain Regions. *PLoS One.* 2016;11(4):e0152342. doi:10.1371/journal.pone.0152342
124. Iwakiri M, Mizukami K, Ikonovic MD, et al. Changes in hippocampal GABABR1 subunit expression in Alzheimer's patients: association with Braak staging. *Acta Neuropathol.* May 2005;109(5):467-74. doi:10.1007/s00401-005-0985-9
125. Shang Y, Das S, Rabold R, Sham JS, Mitzner W, Tang WY. Epigenetic alterations by DNA methylation in house dust mite-induced airway hyperresponsiveness. *Am J Respir Cell Mol Biol.* Aug 2013;49(2):279-87. doi:10.1165/rcmb.2012-0403OC
126. Oddo S. The role of mTOR signaling in Alzheimer disease. *Front Biosci (Schol Ed).* Jan 1 2012;4(3):941-52. doi:10.2741/s310
127. Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology.* Mar 2010;21(2):187-94. doi:10.1097/EDE.0b013e3181cc86e8

128. Alemany S, Crous-Bou M, Vilor-Tejedor N, et al. Associations between air pollution and biomarkers of Alzheimer's disease in cognitively unimpaired individuals. *Environment International*. Dec 2021;157doi:ARTN 106 86410.1016/j.envint.2021.106864
129. Bakulski KM, Seo YA, Hickman RC, et al. Heavy Metals Exposure and Alzheimer's Disease and Related Dementias. *J Alzheimers Dis*. 2020;76(4):1215-1242. doi:10.3233/JAD-200282
130. Murat K, Gruning B, Poterlowicz PW, Westgate G, Tobin DJ, Poterlowicz K. Ewastools: Infinium Human Methylation BeadChip pipeline for population epigenetics integrated into Galaxy. *Gigascience*. May 1 2020;9(5)doi:10.1093/gigascience/giaa049
131. Cristaldi A, Fiore M, Oliveri Conti G, et al. Possible association between PM2.5 and neurodegenerative diseases: A systematic review. *Environ Res*. May 15 2022;208:112581. doi:10.1016/j.envres.2021.112581
132. Fann N, Lamson AD, Anenberg SC, Wesson K, Risley D, Hubbell BJ. Estimating the national public health burden associated with exposure to ambient PM2.5 and ozone. *Risk Anal*. Jan 2012;32(1):81-95. doi:10.1111/j.1539-6924.2011.01630.x
133. Feng S, Gao D, Liao F, Zhou F, Wang X. The health effects of ambient PM2.5 and potential mechanisms. *Ecotoxicol Environ Saf*. Jun 2016;128:67-74. doi:10.1016/j.ecoenv.2016.01.030
134. World Health Organization. *Manual of the international statistical classification of diseases, injuries, and causes of death: based on the recommendations of the ninth revision conference, 1975, and adopted by the Twenty-ninth World Health Assembly*. World Health Organization; 1977.
135. Perin J, Mulick A, Yeung D, et al. Global, regional, and national causes of under-5 mortality in 2000-19: an updated systematic analysis with implications for the Sustainable Development Goals. *Lancet Child Adolesc Health*. Feb 2022;6(2):106-115. doi:10.1016/S2352-4642(21)00311-4

136. Luu TM, Rehman Mian MO, Nuyt AM. Long-Term Impact of Preterm Birth: Neurodevelopmental and Physical Health Outcomes. *Clin Perinatol*. Jun 2017;44(2):305-314. doi:10.1016/j.clp.2017.01.003
137. Ghosh R, Causey K, Burkart K, Wozniak S, Cohen A, Brauer M. Ambient and household PM<sub>2.5</sub> pollution and adverse perinatal outcomes: A meta-regression and analysis of attributable global burden for 204 countries and territories. *PLoS Med*. Sep 2021;18(9):e1003718. doi:10.1371/journal.pmed.1003718
138. Purisch SE, Gyamfi-Bannerman C. Epidemiology of preterm birth. *Semin Perinatol*. Nov 2017;41(7):387-391. doi:10.1053/j.semperi.2017.07.009
139. Woodruff TJ, Parker JD, Kyle AD, Schoendorf KC. Disparities in exposure to air pollution during pregnancy. *Environ Health Perspect*. Jun 2003;111(7):942-6. doi:10.1289/ehp.5317
140. Bonevski B, Randell M, Paul C, et al. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Med Res Methodol*. Mar 25 2014;14:42. doi:10.1186/1471-2288-14-42
141. Yan Q, Liew Z, Uppal K, et al. Maternal serum metabolome and traffic-related air pollution exposure in pregnancy. *Environ Int*. Sep 2019;130:104872. doi:10.1016/j.envint.2019.05.066
142. Liang D, Moutinho JL, Golan R, et al. Use of high-resolution metabolomics for the identification of metabolic signals associated with traffic-related air pollution. *Environ Int*. Nov 2018;120:145-154. doi:10.1016/j.envint.2018.07.044
143. Li Z, Liang D, Ye D, et al. Application of high-resolution metabolomics to identify biological pathways perturbed by traffic-related air pollution. *Environ Res*. Feb 2021;193:110506. doi:10.1016/j.envres.2020.110506
144. Carter RA, Pan K, Harville EW, McRitchie S, Sumner S. Metabolomics to reveal biomarkers and pathways of preterm birth: a systematic review and epidemiologic perspective. *Metabolomics*. Sep 10 2019;15(9):124. doi:10.1007/s11306-019-1587-1

145. Tan Y, Barr DB, Ryan PB, et al. High-resolution metabolomics of exposure to tobacco smoke during pregnancy and adverse birth outcomes in the Atlanta African American maternal-child cohort. *Environ Pollut.* Jan 1 2022;292(Pt A):118361. doi:10.1016/j.envpol.2021.118361
146. Gaskins AJ, Tang Z, Hood RB, et al. Periconception air pollution, metabolomic biomarkers, and fertility among women undergoing assisted reproduction. *Environ Int.* Oct 2021;155:106666. doi:10.1016/j.envint.2021.106666
147. Corwin EJ, Hogue CJ, Pearce B, et al. Protocol for the Emory University African American Vaginal, Oral, and Gut Microbiome in Pregnancy Cohort Study. *BMC Pregnancy Childbirth.* Jun 1 2017;17(1):161. doi:10.1186/s12884-017-1357-x
148. Brennan PA, Dunlop AL, Smith AK, Kramer M, Mulle J, Corwin EJ. Protocol for the Emory University African American maternal stress and infant gut microbiome cohort study. *BMC Pediatr.* Jul 22 2019;19(1):246. doi:10.1186/s12887-019-1630-4
149. Di Q, Amini H, Shi L, et al. An ensemble-based model of PM<sub>2.5</sub> concentration across the contiguous United States with high spatiotemporal resolution. *Environ Int.* Sep 2019;130:104909. doi:10.1016/j.envint.2019.104909
150. Eick SM, Barr DB, Brennan PA, et al. Per- and polyfluoroalkyl substances and psychosocial stressors have a joint effect on adverse pregnancy outcomes in the Atlanta African American Maternal-Child cohort. *Sci Total Environ.* Jan 20 2023;857(Pt 2):159450. doi:10.1016/j.scitotenv.2022.159450
151. Chang CJ, Barr DB, Ryan PB, et al. Per- and polyfluoroalkyl substance (PFAS) exposure, maternal metabolomic perturbation, and fetal growth in African American women: A meet-in-the-middle approach. *Environ Int.* Jan 2022;158:106964. doi:10.1016/j.envint.2021.106964
152. Yu T, Park Y, Johnson JM, Jones DP. apLCMS--adaptive processing of high-resolution LC/MS data. *Bioinformatics.* Aug 1 2009;25(15):1930-6. doi:10.1093/bioinformatics/btp291



153. Uppal K, Soltow QA, Strobel FH, et al. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics*. Jan 16 2013;14:15. doi:10.1186/1471-2105-14-15
154. Wei R, Wang J, Su M, et al. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci Rep*. Jan 12 2018;8(1):663. doi:10.1038/s41598-017-19120-0
155. Mustillo S, Kwon SJTJoMS. Auxiliary variables in multiple imputation when data are missing not at random. 2015;39(2):73-91.
156. Salmon M. riem: Accesses Weather Data from the Iowa Environment Mesonet. 2022.
157. VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. *Epidemiol Methods*. Jan 2014;2(1):95-115. doi:10.1515/em-2012-0010
158. Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of educational and behavioral statistics*. 2002;27(1):77-83.
159. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods*. Aug 2008;40(3):879-91. doi:10.3758/brm.40.3.879
160. Tian L, Li Z, Ma G, et al. Metapone: a Bioconductor package for joint pathway testing for untargeted metabolomics data. *Bioinformatics*. Jul 11 2022;38(14):3662-3664. doi:10.1093/bioinformatics/btac364
161. Uppal K, Walker DI, Jones DP. xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Anal Chem*. Jan 17 2017;89(2):1063-1067. doi:10.1021/acs.analchem.6b01214
162. Schymanski EL, Jeon J, Gulde R, et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol*. Feb 18 2014;48(4):2097-8. doi:10.1021/es5002105

163. Kupsco A, Gonzalez G, Baker BH, et al. Associations of smoking and air pollution with peripheral blood RNA N(6)-methyladenosine in the Beijing truck driver air pollution study. *Environ Int.* Nov 2020;144:106021. doi:10.1016/j.envint.2020.106021
164. Benjamin S, Masai E, Kamimura N, Takahashi K, Anderson RC, Faisal PA. Phthalates impact human health: Epidemiological evidences and plausible mechanism of action. *J Hazard Mater.* Oct 15 2017;340:360-383. doi:10.1016/j.jhazmat.2017.06.036
165. Oaks BM, Adu-Afarwuah S, Ashorn P, et al. Increased risk of preterm delivery with high cortisol during pregnancy is modified by fetal sex: a cohort study. *BMC Pregnancy Childbirth.* Sep 23 2022;22(1):727. doi:10.1186/s12884-022-05061-8
166. Du Y, Hou L, Chu C, Jin Y, Sun W, Zhang R. Characterization of serum metabolites as biomarkers of carbon black nanoparticles-induced subchronic toxicity in rats by hybrid triple quadrupole time-of-flight mass spectrometry with non-targeted metabolomics strategy. *Toxicology.* Oct 1 2019;426:152268. doi:10.1016/j.tox.2019.152268
167. Wei D, Li S, Zhang L, et al. Long-term exposure to PM(1) and PM(2.5) is associated with serum cortisone level and meat intake plays a moderation role. *Ecotoxicol Environ Saf.* Jun 1 2021;215:112133. doi:10.1016/j.ecoenv.2021.112133
168. Patil AS, Gaikwad NW, Grotegut CA, Dowden SD, Haas DM. Alterations in endogenous progesterone metabolism associated with spontaneous very preterm delivery. *Hum Reprod Open.* 2020;2020(2):hoaa007. doi:10.1093/hropen/hoaa007
169. Hu X, Yan M, He L, et al. Associations between time-weighted personal air pollution exposure and amino acid metabolism in healthy adults. *Environ Int.* Nov 2021;156:106623. doi:10.1016/j.envint.2021.106623
170. Weckman AM, McDonald CR, Baxter JB, Fawzi WW, Conroy AL, Kain KC. Perspective: L-arginine and L-citrulline Supplementation in Pregnancy: A Potential Strategy to Improve Birth Outcomes in Low-Resource Settings. *Adv Nutr.* Sep 1 2019;10(5):765-777. doi:10.1093/advances/nmz015

171. Colicino E, Cowell W, Foppa Pedretti N, et al. Maternal steroids during pregnancy and their associations with ambient air pollution and temperature during preconception and early gestational periods. *Environ Int.* Jul 2022;165:107320. doi:10.1016/j.envint.2022.107320
172. Liang L, Rasmussen MH, Piening B, et al. Metabolic Dynamics and Prediction of Gestational Age and Time to Delivery in Pregnant Women. *Cell.* Jun 25 2020;181(7):1680-1692 e15. doi:10.1016/j.cell.2020.05.002
173. Sovio U, Goulding N, McBride N, et al. A maternal serum metabolite ratio predicts fetal growth restriction at term. *Nat Med.* Mar 2020;26(3):348-353. doi:10.1038/s41591-020-0804-9
174. Hood RB, Liang D, Tang Z, et al. Length of PM(2.5) exposure and alterations in the serum metabolome among women undergoing infertility treatment. *Environ Epidemiol.* Feb 2022;6(1):e191. doi:10.1097/EE9.0000000000000191
175. Liang D, Ladva CN, Golan R, et al. Perturbations of the arginine metabolome following exposures to traffic-related air pollution in a panel of commuters with and without asthma. *Environ Int.* Jun 2019;127:503-513. doi:10.1016/j.envint.2019.04.003
176. Zhao L, Fang J, Tang S, et al. PM2.5 and Serum Metabolome and Insulin Resistance, Potential Mediation by the Gut Microbiome: A Population-Based Panel Study of Older Adults in China. *Environ Health Perspect.* Feb 2022;130(2):27007. doi:10.1289/EHP9688
177. Walker DI, Lane KJ, Liu K, et al. Metabolomic assessment of exposure to near-highway ultrafine particles. *J Expo Sci Environ Epidemiol.* Jun 2019;29(4):469-483. doi:10.1038/s41370-018-0102-5
178. Liao J, Gheissari R, Thomas DC, et al. Transcriptomic and metabolomic associations with exposures to air pollutants among young adults with childhood asthma history. *Environ Pollut.* Apr 15 2022;299:118903. doi:10.1016/j.envpol.2022.118903
179. Tang Z, Sarnat JA, Weber RJ, et al. The Oxidative Potential of Fine Particulate Matter and Biological Perturbations in Human Plasma and Saliva Metabolome. *Environ Sci Technol.* Jun 7 2022;56(11):7350-7361. doi:10.1021/acs.est.1c04915

180. Walker DI, Hart JE, Patel CJ, et al. Integrated molecular response of exposure to traffic-related pollutants in the US trucking industry. *Environ Int.* Jan 2022;158:106957. doi:10.1016/j.envint.2021.106957
181. Fenech M. Folate (vitamin B9) and vitamin B12 and their function in the maintenance of nuclear and mitochondrial genome integrity. *Mutat Res.* May 1 2012;733(1-2):21-33. doi:10.1016/j.mrfmmm.2011.11.003
182. Bailey LB, Gregory JF, 3rd. Folate metabolism and requirements. *J Nutr.* Apr 1999;129(4):779-82. doi:10.1093/jn/129.4.779
183. Manta-Vogli PD, Schulpis KH, Dotsikas Y, Loukas YL. The significant role of amino acids during pregnancy: nutritional support. *J Matern Fetal Neonatal Med.* Jan 2020;33(2):334-340. doi:10.1080/14767058.2018.1489795
184. Kalhan SC. One carbon metabolism in pregnancy: Impact on maternal, fetal and neonatal health. *Mol Cell Endocrinol.* Nov 5 2016;435:48-60. doi:10.1016/j.mce.2016.06.006
185. Kalhan SC, Gruca LL, Parimi PS, O'Brien A, Dierker L, Burkett E. Serine metabolism in human pregnancy. *Am J Physiol Endocrinol Metab.* Apr 2003;284(4):E733-40. doi:10.1152/ajpendo.00167.2002
186. Maitre L, Fthenou E, Athersuch T, et al. Urinary metabolic profiles in early pregnancy are associated with preterm birth and fetal growth restriction in the Rhea mother-child cohort study. *BMC Med.* Jul 11 2014;12:110. doi:10.1186/1741-7015-12-110
187. Brosnan JT, Plumptre L, Brosnan ME, et al. Formate concentrations in maternal plasma during pregnancy and in cord blood in a cohort of pregnant Canadian women: relations to genetic polymorphisms and plasma metabolites. *Am J Clin Nutr.* Nov 1 2019;110(5):1131-1137. doi:10.1093/ajcn/nqz152
188. Gaskins AJ, Minguez-Alarcon L, Fong KC, et al. Supplemental Folate and the Relationship Between Traffic-Related Air Pollution and Livebirth Among Women Undergoing Assisted Reproduction. *Am J Epidemiol.* Sep 1 2019;188(9):1595-1604. doi:10.1093/aje/kwz151

189. Go YM, Walker DI, Liang Y, et al. Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicol Sci.* Dec 2015;148(2):531-43. doi:10.1093/toxsci/kfv198
190. Cristaldi A, Fiore M, Oliveri Conti G, et al. Possible association between PM(2.5) and neurodegenerative diseases: A systematic review. *Environ Res.* May 15 2022;208:112581. doi:10.1016/j.envres.2021.112581
191. Henderson J, Carson C, Redshaw M. Impact of preterm birth on maternal well-being and women's perceptions of their baby: a population-based survey. *Bmj Open.* 2016;6(10)doi:ARTN e012676 10.1136/bmjopen-2016-012676
192. Ghosh R, Causey K, Burkart K, Wozniak S, Cohen A, Brauer M. Ambient and household PM2.5 pollution and adverse perinatal outcomes: A meta-regression and analysis of attributable global burden for 204 countries and territories (vol 18, e1003718, 2021). *Plos Med.* Nov 2021;18(11)doi:ARTN e1003852 10.1371/journal.pmed.1003852
193. Bonevski B, Randell M, Paul C, et al. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *Bmc Med Res Methodol.* Mar 25 2014;14doi:ArtN 42 10.1186/1471-2288-14-42
194. Vlaanderen J, Moore LE, Smith MT, et al. Application of OMICS technologies in occupational and environmental health research; current status and projections. *Occup Environ Med.* Feb 2010;67(2):136-43. doi:10.1136/oem.2008.042788
195. Kim S, Hollinger H, Radke EG. 'Omics in environmental epidemiological studies of chemical exposures: A systematic evidence map. *Environ Int.* Jun 2022;164:107243. doi:10.1016/j.envint.2022.107243
196. Maitre L, Robinson O, Martinez D, et al. Urine Metabolic Signatures of Multiple Environmental Pollutants in Pregnant Women: An Exposome Approach. *Environ Sci Technol.* Nov 20 2018;52(22):13469-13480. doi:10.1021/acs.est.8b02215

197. Park J, Kim WJ, Kim J, et al. Prenatal Exposure to Traffic-Related Air Pollution and the DNA Methylation in Cord Blood Cells: MOCEH Study. *Int J Environ Res Public Health*. Mar 10 2022;19(6)doi:10.3390/ijerph19063292
198. Starling AP, Wood C, Liu C, et al. Ambient air pollution during pregnancy and DNA methylation in umbilical cord blood, with potential mediation of associations with infant adiposity: The Healthy Start study. *Environ Res*. Nov 2022;214(Pt 1):113881. doi:10.1016/j.envres.2022.113881
199. Wang C, Plusquin M, Ghantous A, et al. DNA methylation of insulin-like growth factor 2 and H19 cluster in cord blood and prenatal air pollution exposure to fine particulate matter. *Environ Health*. Dec 7 2020;19(1):129. doi:10.1186/s12940-020-00677-9
200. Ghazi T, Naidoo P, Naidoo RN, Chuturgoon AA. Prenatal Air Pollution Exposure and Placental DNA Methylation Changes: Implications on Fetal Development and Future Disease Susceptibility. *Cells*. Nov 5 2021;10(11)doi:10.3390/cells10113025
201. Zhao Y, Wang P, Zhou Y, et al. Prenatal fine particulate matter exposure, placental DNA methylation changes, and fetal growth. *Environ Int*. Feb 2021;147:106313. doi:10.1016/j.envint.2020.106313
202. Yang SI, Lee SH, Lee SY, et al. Prenatal PM(2.5) exposure and vitamin D-associated early persistent atopic dermatitis via placental methylation. *Ann Allergy Asthma Immunol*. Dec 2020;125(6):665-673 e1. doi:10.1016/j.anai.2020.09.008
203. Cai J, Zhao Y, Liu P, et al. Exposure to particulate air pollution during early pregnancy is associated with placental DNA methylation. *Sci Total Environ*. Dec 31 2017;607-608:1103-1108. doi:10.1016/j.scitotenv.2017.07.029
204. Park B, Khanam R, Vinayachandran V, Baqui AH, London SJ, Biswal S. Epigenetic biomarkers and preterm birth. *Environ Epigenet*. Jan 2020;6(1):dvaa005. doi:10.1093/eep/dvaa005

205. Hong X, Sherwood B, Ladd-Acosta C, et al. Genome-wide DNA methylation associations with spontaneous preterm birth in US blacks: findings in maternal and cord blood samples. *Epigenetics*. 2018;13(2):163-172. doi:10.1080/15592294.2017.1287654
206. Baghel R, Maan K, Haritwal T, Rana P. Integration of epigenomics and metabolomics: From biomarkers discovery to personalized medicine. *Epigenetics and Metabolomics*. Elsevier; 2021:31-73.
207. Huang Y, Hui Q, Walker DI, et al. Untargeted metabolomics reveals multiple metabolites influencing smoking-related DNA methylation. *Epigenomics*. Apr 1 2018;10(4):379-393. doi:10.2217/epi-2017-0101
208. Xia Y, Niu Y, Cai J, et al. Effects of Personal Short-Term Exposure to Ambient Ozone on Blood Pressure and Vascular Endothelial Function: A Mechanistic Study Based on DNA Methylation and Metabolomics. *Environ Sci Technol*. Nov 6 2018;52(21):12774-12782. doi:10.1021/acs.est.8b03044
209. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genom Data*. Sep 2016;9:22-4. doi:10.1016/j.gdata.2016.05.012
210. Salas L, Koestler D. FlowSorted.Blood.EPIC: Illumina EPIC data on immunomagnetic sorted peripheral adult blood cells. *R Package Version*. 2022;2.2.0(0)
211. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*. Feb 13 2017;18(1):105. doi:10.1186/s12859-017-1511-5
212. Taibl KR, Liang D, Dunlop AL, et al. Pregnancy-related hemodynamic biomarkers in relation to trimester-specific maternal per - and polyfluoroalkyl substances exposures and adverse birth outcomes. *Environ Pollut*. Apr 15 2023;323:121331. doi:10.1016/j.envpol.2023.121331

213. Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics*. May 1 2012;28(9):1280-1.  
doi:10.1093/bioinformatics/bts124
214. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*. 2014;34(5):502-508.
215. Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. Nov 2017;13(11):e1005752.  
doi:10.1371/journal.pcbi.1005752
216. González I, Déjean S, Martin PG, Baccini A. CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*. 2008;23(12):1-14.
217. Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. Jun 20 2018;14(6):e8124.  
doi:10.15252/msb.20178124
218. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. Mediation: R package for causal mediation analysis. 2014;
219. Hansen K. IlluminaHumanMethylationEPICanno.ilm10b2.hg19: annotation for Illumina's EPIC methylation arrays. *R package version 06.0*. 2017;
220. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. Jan 1 2000;28(1):27-30. doi:10.1093/nar/28.1.27
221. U.S. Census Bureau. Educational Attainment in the United States: 2018. Updated 2021/10/08. Accessed 2023/03/02, 2023.  
<https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html>
222. Maity A, Das B. N6-methyladenosine modification in mRNA: machinery, function and implications for health and diseases. *FEBS J*. May 2016;283(9):1607-30.  
doi:10.1111/febs.13614



223. Zhu X, Fu H, Sun J, Xu Q. Interaction between N6-methyladenosine (m6A) modification and environmental chemical-induced diseases in various organ systems. *Chemico-Biological Interactions*. 2023:110376.
224. Cayir A, Barrow TM, Guo L, Byun HM. Exposure to environmental toxicants reduces global N6-methyladenosine RNA methylation and alters expression of RNA methylation modulator genes. *Environ Res*. Aug 2019;175:228-234. doi:10.1016/j.envres.2019.05.011
225. Li Z, Li N, Guo C, et al. The global DNA and RNA methylation and their reversal in lung under different concentration exposure of ambient air particulate matter in mice. *Ecotoxicol Environ Saf*. May 15 2019;172:396-402. doi:10.1016/j.ecoenv.2019.01.111
226. Gonzalez-Rivera JC, Baldrige KC, Wang DS, et al. Post-transcriptional air pollution oxidation to the cholesterol biosynthesis pathway promotes pulmonary stress phenotypes. *Commun Biol*. Jul 22 2020;3(1):392. doi:10.1038/s42003-020-01118-6
227. Bastonini E, Verdone L, Morrone S, et al. Transcriptional modulation of a human monocytic cell line exposed to PM(10) from an urban area. *Environ Res*. Aug 2011;111(6):765-74. doi:10.1016/j.envres.2011.06.005
228. Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated Omics: Tools, Advances, and Future Approaches. *J Mol Endocrinol*. Jul 13 2018;doi:10.1530/JME-18-0055
229. Larkin A, Hystad P. Towards Personal Exposures: How Technology Is Changing Air Pollution and Health Research. *Curr Environ Health Rep*. Dec 2017;4(4):463-471. doi:10.1007/s40572-017-0163-y