

# A statistical workflow for analyzing the untargeted chemical exposome and metabolome in epidemiologic studies using high-dimensional mixture methods

Anna S. Young, PhD<sup>1,\*</sup>, Chris Gennings, PhD<sup>2</sup>, Stephanie M. Eick, PhD<sup>1</sup>, Donghai Liang, PhD<sup>1</sup>, Douglas I. Walker, PhD<sup>1</sup>

<sup>1</sup>Gangarosa Department of Environmental Health, Emory Rollins School of Public Health, Atlanta, GA, United States

<sup>2</sup>Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, NY, NY, United States

\*Corresponding author: Anna S. Young, PhD, Gangarosa Department of Environmental Health, Emory Rollins School of Public Health, 1518 Clifton Rd NE, Atlanta, GA 30322, United States (ayoung@mail.harvard.edu; anna.s.young@emory.edu).

## Abstract

Humans are exposed to upwards of thousands of chemicals simultaneously, but research has traditionally focused on the health effects of only one chemical at a time. Single-chemical analyses not only underestimate total health risk, but also ignore bias from multicollinearity and co-exposure confounding between chemicals. Advanced statistical mixture methods address these challenges and allow us to both estimate the cumulative health effect of chemical mixtures and identify the strongest chemical contributors. At the same time, untargeted chemical exposome profiling through high-resolution mass spectrometry (HRMS) now supports measurement of over 100,000 chemical signals in biospecimens. However, most mixture methods cannot evaluate untargeted exposome data containing more chemical variables than samples. Weighted quantile sum regression with its recent random subsets implementation (WQS<sub>RS</sub>) is a unique, statistically powerful mixture method for high-dimensional exposome data. It estimates weights of chemicals towards the mixture index over many different repetitions in which only a small random subset of chemicals is used at a time, thus de-correlating data and avoiding overfitting. In this paper, we discuss our statistical workflow and important considerations for the application of WQS<sub>RS</sub> to exposome epidemiology, including manual quantization for non-detects, custom repeated holdouts for matched data, pre-selection of exogenous chemicals, parameter decisions, interpretation options, and visualizations. We then describe its application to functional pathway enrichment analysis with integrated exposome-metabolome data to explore underlying biological mechanisms. These data science approaches will enable exposome epidemiology to discover previously unknown risk factors, estimate cumulative health risk from total chemical mixtures, and gain mechanistic insight.

**Key words:** exposomics; mixture models; multi-omics; high-dimensional data; data science; cumulative chemical mixtures.

## Introduction

### Our chemical ‘soup’ of exposures

Research has traditionally focused on the health impacts of one chemical at a time. In reality, individuals are simultaneously exposed to potentially thousands of environmental chemicals, each of which can add to the cumulative health burden. Based on chemical inventories around the world, over 355,000 chemicals or chemical mixtures have been registered for production and use, including 69,000 in the previous decade alone (2010–2019).<sup>1</sup> For roughly 15% of those registered substances, descriptive chemical names are not provided publicly due to confidential business information, and another ~15% are only ambiguously described.<sup>1</sup> Considering plastic production alone, more than 16,000 chemicals are used or present in products, of which over 4,200 are potentially hazardous and over 10,000 lack hazard information.<sup>2,3</sup> As a result, humans are exposed to complex mixtures of both known and unidentified chemicals, a large

fraction of which have not been comprehensively studied for toxicological safety or health.

A key challenge in identifying new chemical exposures has been the phenomenon of regrettable substitution. Even when one toxic chemical is eventually phased out of production, another chemical—often with a similar structure and same chemical class—may replace it and only later be revealed to have toxicity concerns as well. Such chemical ‘whack-a-mole’ has been observed countless times for phenols in plastic,<sup>4</sup> phthalates in plastic,<sup>5</sup> flame retardants in furniture and electronics,<sup>6</sup> and per- and polyfluoroalkyl substances (PFAS) in consumer products<sup>7</sup>—of which there exist at least 14,000 types of PFAS.<sup>8</sup> Consequently, traditional targeted research methods that measure only a limited number of pre-selected known chemicals cannot match the pace at which new chemicals enter commerce. For example, large human biomonitoring programs have usually analyzed up to about 300 targeted chemicals, due to

Received: January 14, 2025; Revised: July 23, 2025; Accepted: September 24, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

limitations from available analytical standards, cost, and time.<sup>9,10</sup>

## The untargeted chemical exposome

The exposome was first conceptualized in 2005 in response to the evident gulf in the scale at which genomic versus environmental risk factors can be characterized in biospecimens and the need to advance methodologies for measuring biomarkers of environmental exposure at a similar omics level.<sup>11</sup> The most recent definition of exposomics is the study of “the comprehensive and cumulative effects of physical, chemical, biological, and psychosocial influences that impact biological systems by integrating data from a variety of interdisciplinary methodologies and streams to enable discovery-based analysis of environmental influences on health.”<sup>12</sup> As a key aspect of the exposome, untargeted chemical profiling using high-resolution mass spectrometry (HRMS) now supports the simultaneous measurement and characterization of over 100,000 chemical signals in human biospecimens, including both the internal exposome (ie, exogenous environmental chemicals and their biotransformation products) and the metabolome (ie, endogenous metabolites).<sup>13-15</sup>

The untargeted approach encompasses even unknown chemicals that we cannot yet identify but can still monitor in samples and elucidate structural information for.<sup>16</sup> Each untargeted chemical feature is characterized in the mass spectrometry output by its accurate mass-to-charge ratio ( $m/z$ ) and retention time.<sup>17</sup> Although the data generated are relative abundances (ie, ion intensities) rather than absolute targeted concentrations,<sup>18</sup> HRMS offers a powerful hypothesis-free *discovery* approach for detecting previously unknown environmental risk factors of disease as well as the metabolic responses that underlie this risk, without requiring analytical standards for the chemicals.<sup>19,20</sup> With the high dimensionality of untargeted chemical data produced, statistical analysis remains a major challenge in the field of exposomics. The objective of our commentary is to share a statistical workflow that we have developed for evaluating the cumulative mixture effects of untargeted chemicals on disease, prioritizing novel chemicals of concern, and exploring mechanisms of action. This workflow has been applied in our recent studies of exposome drivers of fertility outcomes<sup>21</sup> and lymphoma risk.

## The importance of statistical mixture methods

In response to the need to evaluate exposures to chemical and non-chemical stressors as complex mixtures, there has been a recent explosion in statistical mixture methods.<sup>22</sup> For purposes of this paper, we focus on the chemical exposome and consider a “mixture” to be the cumulative (combined) exposures to multiple chemicals through the same and/or different routes of exposure.<sup>23</sup> Mixture methods address two key statistical limitations of the traditional single-chemical regression approach that models the effects of one chemical exposure at a time.

First, single-chemical regression may underestimate risk, because many chemicals can simultaneously affect the same health endpoint or receptor (through similar or different mechanisms) and thus accumulate the health burden. A chemical may even have a small effect size that is not statistically observable on its own, so its hazard would go overlooked and unaddressed in the absence of assessments of cumulative mixture effects.<sup>24</sup> Some chemicals actually interact together in non-additive ways

to amplify, trigger, or attenuate each other's effect;<sup>23</sup> however, not all mixture methods can account for complex interactions.

Second, and most importantly, traditional regression ignores bias arising from collinearity or confounding due to co-exposure of other chemicals. The issue arises from the high degree of correlation between individual chemicals, between or across chemical classes, and between chemical metabolites. For example, an analysis of typical chemical exposure data from the U.S. National Health and Nutrition Examination Survey (NHANES) found 2,656 significant pairwise correlations among 289 exposure variables, demonstrating a potentially dense correlation structure.<sup>25</sup> A pair of chemicals has the potential to confound each other<sup>26</sup> if they are highly correlated due to a common source (such as flame retardants in furniture),<sup>27</sup> common route of exposure (such as PFAS in drinking water),<sup>28</sup> and/or common biotransformation pattern (such as multiple urinary metabolites of the same parent phthalate chemical).<sup>29</sup> For example, in a single-chemical regression model, this confounding may result in Chemical B incorrectly appearing to be associated with the outcome (through an open backdoor path on a Directed Acyclic Graph [DAG]), when in reality Chemical A is the true risk factor. In that specific case, mutually adjusting for both chemicals within the same regression model would yield the correct result. However, in other situations, such as in the presence of an unmeasured confounder of Chemical A, mutual adjustment of both chemicals can worsen the bias for Chemical A's association, or even reverse the direction of the association for Chemical B due to conditioning on a collider.<sup>26</sup> The paper by Weisskopf et al. provides helpful DAGs to demonstrate these case scenarios of bias amplification from exposure correlation patterns.<sup>26</sup> Relatedly, the “reversal paradox” refers to the case in multiple-chemical regression models when the coefficients for two highly correlated exposures associated with the outcome reverse direction in opposite extremes from each other, although this does not always happen.<sup>30,31</sup> While the potential bias arises even with only two correlated chemicals in the same regression model, it can grow in complexity when adding more than two or when the correlations strengthen, including 1) the potential for worse bias amplification,<sup>26</sup> 2) more inflated standard errors and less stable estimates due to multicollinearity (since the chemicals are linear predictors of each other),<sup>32-34</sup> and 3) overfitting of the model with more variables than the sample size can handle (such that the model is simply fitting noise).<sup>35</sup> As such, model estimates for highly correlated chemical exposures in traditional multivariate regression are not reliable. Specialized mixture methods seek to address the high dimensional, multicollinear structure of complex chemical exposure data.

## Special considerations for untargeted mixtures

Although statistical mixture methods embrace the complexity of chemical exposures, most approaches cannot handle untargeted data at the omics-scale. The major barrier with untargeted omics data lies in the fact that the number of parameters is far greater than the sample size ( $p \gg n$ ), and thus would overfit the model to an impossible degree. In addition, untargeted data can be even more highly correlated because of the presence of some redundant signals, such as adducts, isotopes, or fragments of the same chemical, and multiple biotransformation products arising from the same exogenous chemical.<sup>36</sup> For this reason, not all mixture methods can be readily scaled up to the high dimensions of the exposome.

## Limitations of other statistical approaches

Choice of mixture method largely depends on the research question.<sup>22</sup> In our case, we aim to use mixture methods to evaluate cumulative mixture effects from *untargeted* chemical exposures and identify the ‘bad actor’ chemicals driving the mixture effect the most. There are several common statistical approaches that would not work best for our research questions.

For example, some studies calculate the sum, molar sum, or potency-weighted sum of specific chemical classes to model the total effect on health using fewer variables.<sup>37</sup> This approach has been commonly applied to phthalates or phthalate subgroups, such as the summed urinary metabolites of di(2-ethylhexyl) phthalate (DEHP) or high versus low molecular weight phthalates.<sup>29</sup> However, chemical sums lose data resolution and can hinder interpretations about individual chemicals for decision-making. In addition, the summation could mask hazards if chemicals in the same class have different molecular weights, abundances, toxicological relevance, interactions, concentrations at which adverse effects occur, or even opposing directions of effects. For instance, the health effect of a low-concentration chemical may not be observed when summed with a high-concentration chemical that does not affect the outcome.<sup>32</sup>

Bayesian kernel machine regression (BKMR) is a popular mixture method<sup>38</sup> that can model the joint effects of exposures using a flexible kernel function that allows for non-linear, smooth exposure-outcome relationships and interactive effects between chemicals (with some sacrifices to statistical power).<sup>39,40</sup> Although useful for targeted chemical mixtures, especially research questions related to non-linearity and interactions between chemicals, BKMR is currently not practical for higher dimensional chemical data ( $p \gg n$ ) because of the large sample size and computational intensity needed for the non-parametric kernel function, which is highly flexible but less statistically powerful.<sup>41</sup> In addition, the visual interpretation of curvilinearity and interaction within high-dimensional mixtures in BKMR would be intractable. For applications of mixture methods to lower-dimensional mixtures, a recent publication offers a helpful workflow for statistical decisions related to the distribution and type of data, variable transformations, missing data, statistical assumptions, specific research questions, and other study design considerations.<sup>42</sup>

Quantile g-computation (QGcomp) is another common mixture method similar to WQS that builds a weighted summary index based on quantiles of exposures, with a few slight advantages or disadvantages depending on the desired research question and sample size. First, QGcomp relaxes the directional homogeneity assumption by incorporating opposing effects within the same index,<sup>43</sup> whereas WQS now sequentially defines separate positive and negative indices to evaluate double associations of the mixture in both directions.<sup>44,45</sup> Second, QGcomp allows for more non-additivity and non-linearity of effects if specified as model terms in advance. Third, QGcomp does not split samples into repeated training/validation sets, which has tradeoffs between statistical power and computational speed versus generalizability and weight stability.<sup>43,46</sup> However, for the purpose of untargeted mixture methods in this paper, QGcomp has not yet been designed for applications to high-dimensional data.<sup>47</sup>

Other statistical approaches seek to systematically reduce the dimensions of data before modeling. For example, principal component analysis (PCA) is an unsupervised approach that converts the chemical data into a smaller set of uncorrelated linear

predictors (not taking into account the outcome). PCA may be helpful for assessing patterns in exposures (eg, shared sources of chemicals) without bias from multicollinearity, however, there is a loss of information, generalizability (due to dimensionless units), and interpretability of the individual principal components that makes it difficult to identify specific ‘bad actor’ chemicals or their doses of harm when modeling the effects on health outcomes.<sup>48</sup> Other dimension reduction approaches that do consider the outcome (ie, are supervised) include shrinkage methods. Ridge regression shrinks the regression coefficients in a manner that decreases variance (with a trade-off of increased bias),<sup>49</sup> but it keeps all predictors in the model and thus does not reduce the dimensions of the data.<sup>31,50</sup> Lasso regression builds on ridge to allow coefficients to shrink to exactly zero, thus producing a model with fewer predictors.<sup>51</sup> However, lasso saturates at  $n$  predictors (no more than the sample size), and it tends to select one exposure arbitrarily from a group of highly correlated exposures,<sup>50</sup> which may erroneously lead the researcher to conclude that the other unselected exposures are not associated with the health outcome.<sup>31</sup> Elastic net combines both lasso and ridge regression and can work in a high-dimensional setting; however, it encourages a “grouping effect” that either keeps all or eliminates all the chemicals in a highly correlated set, even when some of the chemicals in the set may only be correlated due to shared exposure routes not health outcomes.<sup>31</sup> This demonstrated poor specificity in the selection of correlated variables by elastic net compared to WQS is not ideal for studying chemical exposures.<sup>31,52</sup> Adaptive elastic net has been proposed to be more successful than its predecessor in the case of  $p \gg n$ , but it does not provide a cumulative estimate across individual exposures to address research questions focused on combined mixture effects.<sup>53</sup>

## Weighted quantile sum regression with random subsets of untargeted chemicals

Recent extensions of weighted quantile sum (WQS) regression now support estimation of cumulative mixture effects and individual contributions of untargeted chemicals under  $p \gg n$  scenarios.<sup>54</sup> To our knowledge, this is currently the only mixture method that can do so under typical epidemiologic cohort sizes with high-dimensional exposome data without loss of information for identifying individual ‘bad actor’ chemicals.

### WQS method overview

At its base, WQS calculates a mixture effect by combining quantiles of all the exposures (here, chemicals) into one summary weighted index that represents the cumulative mixture effect on the health outcome in a single direction.<sup>31,52</sup> To do so, in a training set of the data (eg, 40% of participants), each chemical is assigned a weight that reflects its individual contribution to the overall mixture effect on the outcome, where all weights sum to one. Every participant then has a score (mixture index) based on this formula for the weighted sum of their personal chemical exposures. In the final regression model using the testing set of the data (eg, the other 60% of participants), WQS simply determines the association of the mixture index with the outcome in a single degree-of-freedom, highly statistically powerful test. Quantiles are used in WQS instead of the fully continuous data because they convert the exposures to all be on the same scale in the mixture regardless of units and avoid extreme values that would cause weights to grow in extremes.<sup>31</sup>

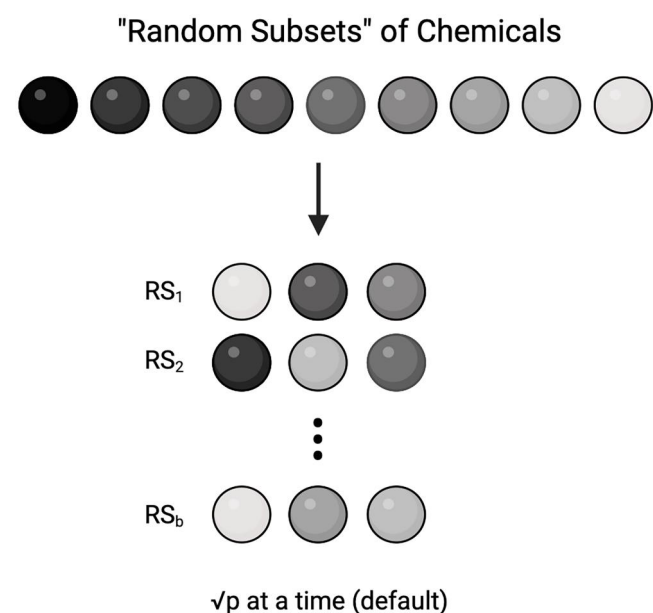
It is best practice to always apply the repeated holdouts (RH) variation of WQS, which randomly partitions the data

observations into the training set (used to estimate weights) and the testing/validation set (used to test the mixture effect) many times over.<sup>46</sup> The entire WQS model is repeated within each of, say, 100 repeated holdouts. This produces a distribution of validated results, which improves data representativeness and generalizability, stabilizes the chemical weights, and characterizes uncertainty in the estimates through confidence intervals.

For the critical extension to untargeted exposome data, the random subsets (RS) implementation of WQS addresses the challenges from having more exposure parameters than sample size ( $p \gg n$ ).<sup>54</sup> In the basic version of WQS, chemical weights are estimated by bootstrapping over observations in the training data (eg, 1,000 randomly selected bootstrap samples with replacement) and then determining the average weights based on the relative signal of the test statistic for the mixture index slope (in association with the outcome) in each bootstrap sample, using the full set of chemicals each time.<sup>31</sup> By contrast, the RS version of WQS uses “feature bagging” to estimate weights over many different randomly selected subsets of chemicals in the full training data (Figure 1). This allows WQS<sub>RS</sub> to aggregate across de-correlated sets of chemicals by repeatedly perturbing the exposure data under different correlation scenarios, thus avoiding multicollinearity and co-confounding, improving generalizability, and preventing overfitting.<sup>54</sup> The RH and RS variations should be used together, which performs the random subsetting procedure within every repeated holdout. Chemical weights are thus estimated a total of RH x RS times, so the computational intensity of WQS<sub>RS</sub> can grow quickly. In summary, WQS<sub>RS</sub> represents a major advancement in mixture methods for high-dimensional exposure data with complex correlation patterns.

### Selecting the chemical mixture input

Although WQS<sub>RS</sub> supports high-dimensional untargeted data, inputting a larger and larger number of chemical features as exposures into the model does not necessarily improve the usefulness of the model. A summary of our roadmap is provided in Figure 2.

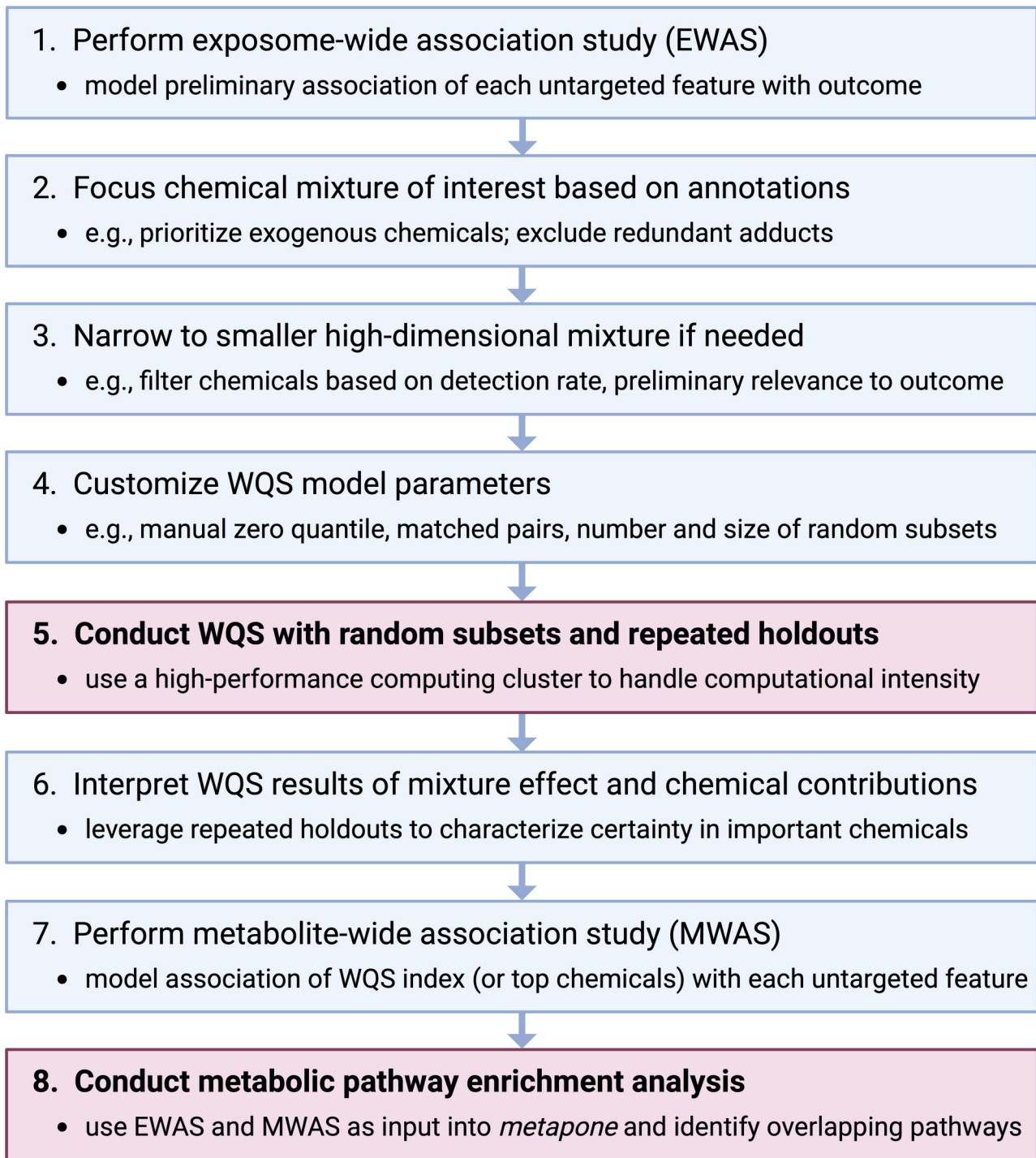


**Figure 1.** Diagram showing how the Random Subsets extension of WQS repeatedly estimates chemical weights over  $b$  different scenarios of smaller chemical mixtures, thus de-correlating the exposure data. For purposes of illustration, only nine chemicals are included in this example.

First, depending on the research question, it may be beneficial to focus the mixture index to only environmental risk factors by excluding possible endogenous or pharmaceutical metabolites and early biomarkers of disease that could overinflate the mixture effect. This can be achieved through a variety of approaches, each of which faces a tradeoff between either excluding more possible endogenous metabolites or including more unidentified chemical features of unknown origin. For example, the chemical mixture could focus on only the detected features with annotations as possible environmental chemicals based on database matches, thus minimizing the presence of endogenous features (whether identified or unidentified) but being less inclusive to novel environmental chemicals not yet existing in databases. The coverage offered by a particular annotation database may need to be pruned depending on the research hypothesis; for example, the Norman Substance Database combines multiple sources of environmental chemical lists, some of which are pharmaceutical-focused or cover drinking water contaminants that include pharmaceuticals and hormone-related compounds.<sup>55</sup> Further, any additional annotated adducts or isotopes of the detected environmental chemicals could be removed to reduce feature redundancy (retaining the primary M + H or M - H adduct), although this should only be done among the annotations meeting the highest confidence level of three by xMSannotator (ie, adducts predictably clustered into the same correlation modules, retention time sub-modules, and mass-defect sub-groups).<sup>56</sup> An alternative approach for environmental chemical selection could focus on including all the chemical features that do not have identities or annotations as possible endogenous features, which would be inclusive to more chemicals but vary in effectiveness depending on the metabolomic database's degree of coverage and endogenous/exogenous distinction. The strictest approach would be to focus the mixture on specific known classes or source groups of environmental chemicals. The potential interference by endogenous biomarkers is more pronounced in data from liquid chromatography (LC) HRMS, which widely integrates both the metabolome and the chemical exposome in its measurement of polar molecules with specific functional groups. Depending on sample preparation and extraction methods, gas chromatography (GC) HRMS can allow more focused detection of environmental chemicals, which is why the use of both instruments has been recommended for optimal chemical coverage.<sup>10,15</sup> In summary, the purpose of the WQS index and what mixture it intends to represent should be carefully considered in advance of modeling.

Second, the more chemicals included, the more repetitions needed, so the computational intensity can grow rapidly. Depending on the number of detected environmental chemicals in the research study, a layered pre-filtering strategy may help. The first layer is to restrict chemicals by some detection rate threshold to ensure there is sufficient spread of exposure across the quantiles for each chemical. Stricter thresholds (such as 50% or 75%) may miss out on important chemicals that are mostly found only in those that develop the disease, so this should be tailored to the study; in our 1:1 matched case-control study, for example, we are using 25% to allow for chemicals that hypothetically only present in half of the cases and none of the controls. However, quantizing data with such low detection thresholds requires a specialized approach (see next section).

As the next pre-filtering criterion, the mixture inputted into WQS<sub>RS</sub> could be narrowed to only exposures with potential health relevance, while being cautiously inclusive. This is because including higher numbers of irrelevant chemicals may



**Figure 2.** An untargeted exposome-metabolome statistical workflow to leverage weighted quantile sum regression (WQS) with random subsets for identifying cumulative mixture effects on a health outcome and determining bad actor chemicals driving the mixture effects, then to use pathway enrichment analysis to investigate metabolic pathways that may underlie the mixture effects.

unexpectedly attenuate the mixture effect if the random subset parameter is not equivalently scaled, despite that the model trends weights of unimportant chemicals close to zero. To illustrate this, in the hypothetical case of a model with 18,000 chemicals compared to a model with 1,800 relevant chemicals, if the number of random subsets of chemicals remains the same (eg, 2,000 repetitions, even with slightly larger subset sizes), there would likely be more random subsets with mixtures full of the irrelevant chemicals that exhibit null effects, which then get

averaged into the overall weighted index. Then, weight estimates may be based more on noise than signal. Instead, the mixture could first be further filtered to only the chemicals univariately associated with the outcome under a loose significance threshold without correcting for multiple testing (such as unadjusted  $p < 0.10$  or even  $p < 0.20$  in the adverse direction) in an exposome-wide association study (EWAS). We err on the side of inclusivity since univariate analyses are prone to issues from multicollinearity or chemical co-confounding and, under stricter significance

thresholds, might miss chemicals that would show a stronger effect in the de-correlated mixture model. Although not without limitations, this pre-filtration approach has advantages over high-dimension shrinkage methods. As we described in the introduction, elastic net regression tends to either keep all or eliminate all chemicals among a highly correlated group,<sup>31</sup> and thus could over-exclude potentially relevant chemicals from further consideration. However, future work should use simulated data to compare different approaches to selecting high-dimensional mixtures.

Finally, a third possible filtering layer is to conduct the WQS<sub>RS</sub> models separately for the data from different instruments (GC-HRMS Versus LC-HRMS) or different instrument columns (LC hydrophilic interaction [HILIC] chromatography versus C18 hydrophobic interaction chromatography), if applicable. This not only reduces potential overlap in chemicals detected through each methodology but can also demonstrate robustness of results across platforms. However, the tradeoff is not identifying the fullest cumulative mixture effect from incorporating all relevant chemicals. Overall, the decisions about the chemical mixture to input into WQS<sub>RS</sub> should depend on the research question and the dimensions of the exposome data in the specific study, and conducting sensitivity analyses under different mixture sizes and model parameters is helpful to understand the consistency of results.

### Important decisions on WQS parameters

For the choice of the number of repeated holdouts of observations to implement, 100 RHs are typical and sufficient to improve generalizability over traditional epidemiologic analyses and to characterize uncertainty of weights in mixture indices.<sup>46</sup> Although more RHs (such as >1,000) would approximate the normal distribution better, the computational requirements could be prohibitive: 1,000 RHs would take 10 times as long as 100 RHs would. By contrast, the choice of the number and size of random subsets of chemicals is more important to customize to each research question. We recommend first trying various choices of RSs under RH = 1 (for lower computational intensity) to see how the distribution of chemical weights change. If the number of chemicals with weights extremely near zero decreases with more RSs, that might mean that some chemicals did not have sufficient chance of being included under the lower number of RSs and thus that a higher number of RSs is still offering benefits. The size of each RS by default is the square root of the number of chemicals included in the mixture, and its input has a trade-off between giving a particular chemical more chances to be included in the model (under a fixed number of RSs) versus better de-correlating the data. In sensitivity analyses in our fertility study, we found the results to be sensitive to the choice of RS size. When we increased the number of chemicals by loosening filter criteria to include potentially less relevant exposures, and the RS size increased automatically, there were fewer important mixture contributors meeting Busgang criteria than when fixing the RSs at the original smaller size.<sup>21</sup> We recommend to choose a relatively small size of each RS that allows for sufficient perturbation of chemical correlation patterns to discover important exposures, assuming that the number of RS repetitions is high enough to still give each chemical enough chance of being included in RSs. The previous WQS simulation study of the random subset implementation used a mixture of 472 untargeted metabolites with 1,000 RSs and a default RS size of 22 (ie, &scenveo\_unknown\_entity\_Symbol\_F0D6;472),<sup>54</sup> but simulation

analyses have not yet been conducted on implications of higher RS sizes due to larger high-dimensional mixtures.

The average weights of chemicals are estimated based on the relative signal of the test statistic for the index slope in each random subset, and this signal function parameter has multiple possible values. For example, an “expt” signal function would apply the exponential of the t-statistic, which allows the most important chemicals to be much higher weighted than others, if that is desired. In addition, the  $\beta$  slopes of the indices can be constrained to a single direction, which excludes random subsets with slopes in the opposite direction and ensures that only those in the relevant direction will contribute to the estimation of weights; thus, the mixture effect represents the adverse direction of harm.

There are two key deviations from default parameters that are important to consider. First, we recommend that the exposure data are manually quantized into deciles or quartiles before input into the WQS model (with the  $q$  parameter then set to null) in such a way that the non-detect values are put into their own zero-quantile and the detected values are quantized separately (eg, into 9 quantiles for a total of 10 deciles). The quantization is done individually for each chemical and modified from the source code for function `gwqs_rank` (see example code at [github.com/anna-s-young/exposome-statistics](https://github.com/anna-s-young/exposome-statistics)). This approach avoids chemicals having, for example, five deciles that all refer to zero values and then large jumps between the latter five deciles. Second, matched case-control studies may need to consider how data are partitioned. The default is to randomly split the observations (eg, participants) into separate training and validation sets for each repeated holdout. However, this may lose some of the benefits of individually matched case-control pairs if a particular set no longer has a similar distribution of confounders between cases and controls. An alternative is to manually partition the pairs, instead of individual observations, randomly within each repeated holdout. This is supported by the “validation\_rows” parameter in the model that takes a list of vectors (one for each RH) indicating the rows of observations to include in the validation set (see code available at [github.com/anna-s-young/exposome-statistics](https://github.com/anna-s-young/exposome-statistics)).

### Computational intensity

The WQS<sub>RS</sub> models with high-dimensional exposome data should be performed on a high-performance computing (HPC) cluster rather than a standard computer due to the computational intensity. The speed can be accelerated through parallel processing by using the “multisession” or “multicore” option for the plan strategy in the WQS function call. Although computational time can vary greatly for a cluster job depending on factors such as network performance, processing speed, cluster utilization, and available memory, in our unpublished lymphoma nested case-control study we completed most WQS sensitivity models in approximately 7-60h each under multisession (using mixtures of up to 2,700 chemicals, up to 4,000 RSs, 100 RHs, 444 samples, and our slower customizations of manual quantization and manual partitions) on the HPC at the Emory Rollins School of Public Health. Without the manual partitioning of case-control pairs into repeated holdouts, the models took an order of magnitude less time. In our fertility study, we also completed most of the sensitivity models in 4-50h each (using mixtures of up to 2,700 chemicals, up to 4,000 RSs, 100 RHs, 82 samples, and manual quantization).<sup>21</sup> Increasing the number of chemicals and/or numbers of RSs to much higher degrees were our primary limiting factors for cluster load. However, we observed in the

sensitivity analyses that the margins of return diminished with increasing numbers of RSs after a certain point, which justified us keeping 2,000 RSs for our main models (although this number will change depending on the study and mixture size). As described earlier, the choice of RSs can first be decided upon while using  $RH = 1$  in faster preliminary models. Importantly, the independence of the repeated holdouts and random subsets makes them easily parallelizable and able to be scaled up in line with available high-performance computing resources.

## Interpreting the WQS results

With repeated holdout validation, the mixture effect ( $\beta$  estimate for the index's association with the outcome) is interpreted as an aggregation across repeated holdouts. The function will provide a mean and median of the mixture effect, along with confidence intervals. The median and percentile-based interval are preferred because they do not make assumptions about symmetry. The units of the mixture effect are per quantile. However, a histogram of the distribution of the cumulative mixture index across observations may show that it bunches in the middle quantiles, as it may be rare for a participant to have consistently low (or consistently high) quantiles for most exposures. If so, the per-quantile mixture effect may represent a difference covering a wide range of the exposures. An alternative is to transform the units to a per standard deviation (SD) change in the index. To do so, the  $\beta$  estimate within each repeated holdout must be manually extracted from the output and multiplied by the standard deviation of that holdout's index, then the median or mean of the estimates is used as the mixture effect per SD increase. This approach may also aid in the comparison of the cumulative mixture effect to the magnitudes of individual chemical effects, which can be estimated (and transformed to per-SD) in basic univariate regression models among the chemicals deemed important. Such a comparison can show the usefulness of  $WQS_{RS}$  for evaluating chemicals as mixtures instead of as single exposures.

The weights estimated for each chemical component can reveal potential 'bad actor' chemicals driving the overall mixture effect. However, because the weights always sum to one, they should only be interpreted in the case when the mixture effect itself is significant or borderline significant, based on its overall  $p$  value or percent of repeated holdouts in which it reaches significance. In the simplest interpretation of weights, chemicals are considered important if they have an average weight higher than the equi-weight threshold ( $1/p$ ), which represents the hypothetical scenario of equal contributions by all chemicals. However, with the repeated holdout validation, we can better characterize the uncertainty in the chemical weights. Following the "Busgang criteria", chemicals with an average weight above the threshold within at least 90% of repeated holdouts can be defined as "probable contributors", else within at least 50% of holdouts as "possible contributors", else within at least 10% of holdouts as "possibly not contributors", else within less than 10% of holdouts as "probably not contributors."<sup>57,58</sup> These criteria characterize how replicable the results are across data samples from the same underlying population, especially because certain chemicals may be misclassified as concerning—or not concerning—when looking at only one holdout of data.<sup>46</sup> We can graph these distributions of weights (eg, among the "possible contributors") using the individual repeated holdout data from the WQS output. An example for chemicals across different levels of contribution is shown in Figure 3 (code available at [github.com/anna-s-young/exposome-statistics](https://github.com/anna-s-young/exposome-statistics)).

## Strengths and limitations of WQS

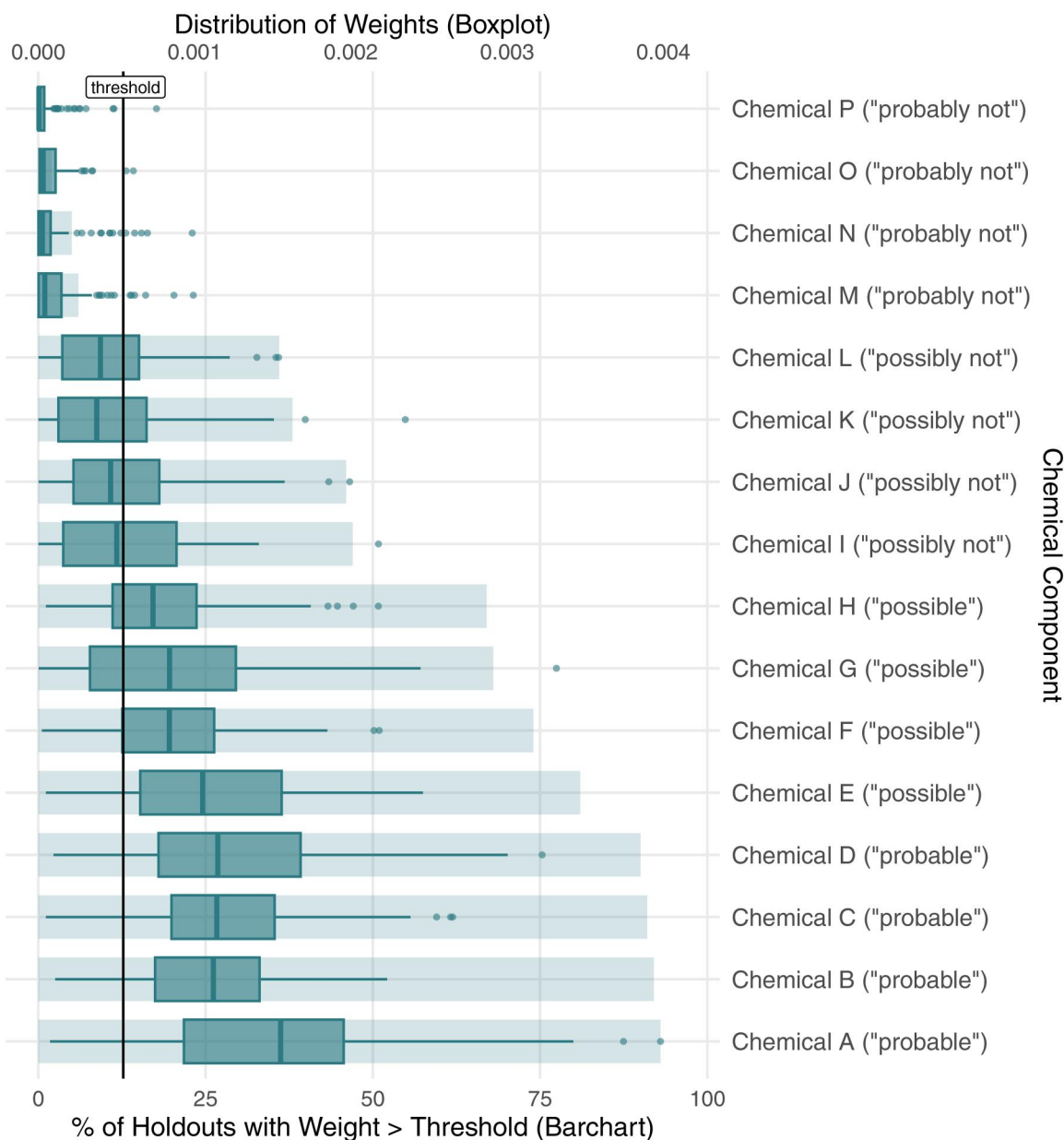
$WQS_{RS}$  is a powerful approach to identify cumulative mixture effects and bad actor chemicals while uniquely embracing the complexity and high dimensionality of untargeted exposome data. To our knowledge, it is currently the only mixture method that can be tailored to epidemiologic studies where the number of exposures exceeds the sample size. Compared to traditional univariate regression analyses, this mixture method avoids bias from multi-collinearity and co-confounding of chemicals and thus more accurately prioritizes chemicals of concern. At the same time,  $WQS_{RS}$  is highly statistically powerful, as it only conducts a single degree-of-freedom test, *without* losing information critical for interpretation of individual chemical risk factors. The summary mixture index not only provides a measure of cumulative mixture effects from the many chemicals that can simultaneously interfere with health, but it can also be used in other analyses as a single variable representing harmful exposure (see next section).

As limitations, WQS assumes that there are no interactions between exposures and that there is a constant change in risk between quantiles. These assumptions can be tested in sensitivity analyses considering other quantiles (eg, quartiles instead of deciles) and in other mixture methods considering small mixtures (such as Bayesian kernel machine regression).<sup>39</sup> With the random subsets implementation, WQS also assumes that all component effects are in the same direction, when in reality, the chosen chemical mixture may include substances that operate in the non-adverse direction, such as endogenous metabolites or some chemicals seeming to be protective due to unknown confounding (eg, pesticides related to diet and nutrition). Or, the health endpoint may be adverse in either extreme direction, as opposed to a simple dichotomy. In some cases, especially under hypothesis discovery, determining the direction of interest is not straightforward. Despite the limitation, this unidirectionality assumption actually helps prevent the reversal paradox arising from complex multicollinearity of exposures,<sup>45</sup> while also focusing the index for better interpretability. There is a recent extension of WQS for double (positive and negative) indices with a penalization term, but it does not yet support random subsets of chemicals for high-dimensional data at the same time.<sup>45</sup> As another limitation that should be acknowledged in matched case-control studies, WQS does not yet allow for conditional logistic regression in its models, only adjustment for the matching variables. WQS currently supports linear, logistic, Poisson, quasi-Poisson, and negative binomial regression in its current version 3.0.5 of the R package *gWQS*. Furthermore, WQS only supports continuous or ordinal exposure variables, however, categorical variables can be used if transformable to an ordinal structure. For example, previous work has evaluated quartiles of scales for post-traumatic stress disorder symptoms, depressive symptoms, and life stressors.<sup>59,60</sup> Although quantization of continuous exposures loses the full range of levels, it prevents extreme weights from outliers.<sup>31</sup> Finally, the mixture method can become quite computationally intensive depending on the number of exposures, random subsets, repeated holdouts, and samples, and it is best used on a high-performance computing cluster.

## Metabolic pathway enrichment with WQS

### Pathway enrichment method overview

Because of the simultaneous measurement of both the chemical exposome and metabolome using untargeted HRMS,<sup>13-15</sup>



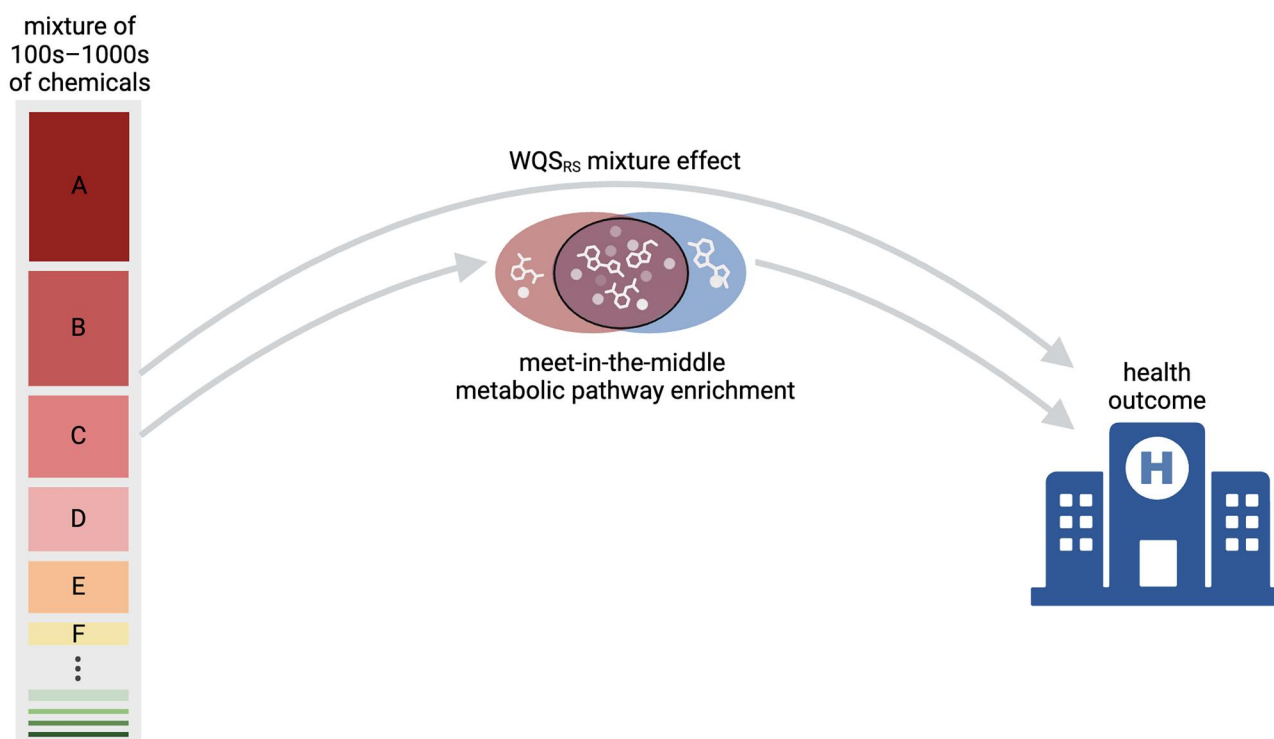
**Figure 3.** Example visualization of distributions of weights of chemicals across repeated holdouts of data in weighted quantile sum regression (WQS) models. The equi-weight threshold for weights is defined as one divided by the number of included chemicals. Here, a random selection of chemicals across different contribution levels are shown.

multi-omics analysis offers the opportunity to identify biological mechanisms that may underly the associations between chemical exposure and adverse health outcomes. For example, a meet-in-the-middle (MITM) strategy could: (1) identify the metabolic pathways associated with disease, (2) identify the metabolic pathways associated with exposure, and then (3) determine which significant pathways overlap “in the middle” between exposure and disease (Figure 4); however, researchers have approached MITM in different ways.<sup>61</sup> Because the untargeted metabolomic data are also high dimensional, the supervised WQS<sub>RS</sub> index is useful as a single exposure variable representing the cumulative mixture effect of untargeted chemicals on disease, thus reducing the complexity of one of the multi-omics layers and focusing the exposure index to the chemicals that are most relevant to the outcome and its mechanisms. Ideally, prospective, longitudinal samples would be used such that the

metabolome data lie chronologically in the middle between the exposome and the health outcome and thus avoid problems with reverse causality. However, achieving this temporality is not always possible due to limited sample availability or budget restrictions, in which case care should be taken to acknowledge the potential for reverse causality or health treatment effects.

To accomplish steps 1 and 2 of MITM (separately), *p* values are first calculated for the univariate associations of each detected untargeted feature with the dependent variable (either the health outcome for step 1 or the exposure for step 2). Then, functional pathway enrichment analysis can predict functional activity by mapping all possible metabolite annotations for each feature (based on *m/z* and/or retention time) to a metabolic network and then finding the significant features that are locally enriched on a structure (ie, represent biological activity), whereas the false matches would only be randomly distributed in the network.<sup>62</sup>





**Figure 4.** Diagram of our approach to assess the mixture effect of the untargeted exposome on health and then analyze underlying metabolic pathways that are significantly enriched for both the WQS mixture index and the health outcome (ie., that overlap in the middle). Note that prospective longitudinal samples are best for interpreting causality. Note: WQS = weighted quantile sum.

This method leverages metabolic interconnections to improve prediction of pathway activity without having to identify each metabolite *a priori*.<sup>62</sup> *Mummichog* (which is also implemented in *MetaboAnalyst*) has been the most common algorithm for pathway enrichment analysis,<sup>62,63</sup> however, we currently use *Metapone*'s permutation-based weighted hypergeometric test for several reasons.<sup>64</sup> *Metapone* jointly analyzes both positive- and negative-ion mode HRMS data when applicable (while avoiding double counting), accounts for matching uncertainty by using fractional counts of features (thus down-weighting those with higher numbers of matches), has an R package for suitability in our workflow alongside environmental mixture methods, and combines pathway information from multiple databases for higher relevance to xenobiotic pathways.<sup>64</sup> It leverages three databases: *mummichog*,<sup>62</sup> the Kyoto Encyclopedia of Genes and Genomes (KEGG),<sup>65</sup> and the Small Molecule Pathway Database (SMPDB).<sup>66</sup> The overall *p* value of enrichment significance for each pathway is defined as the proportion of permutations in which the total fractional count of significant features in that pathway based on the real feature *p* values is lower than the total fractional count of significant features in that pathway based on randomly permuted *p* values for each feature.<sup>64</sup> As the final step in MITM, the results from pathway enrichment for exposure versus outcome can be compared for potential overlap (example in [Figure 5](#)).

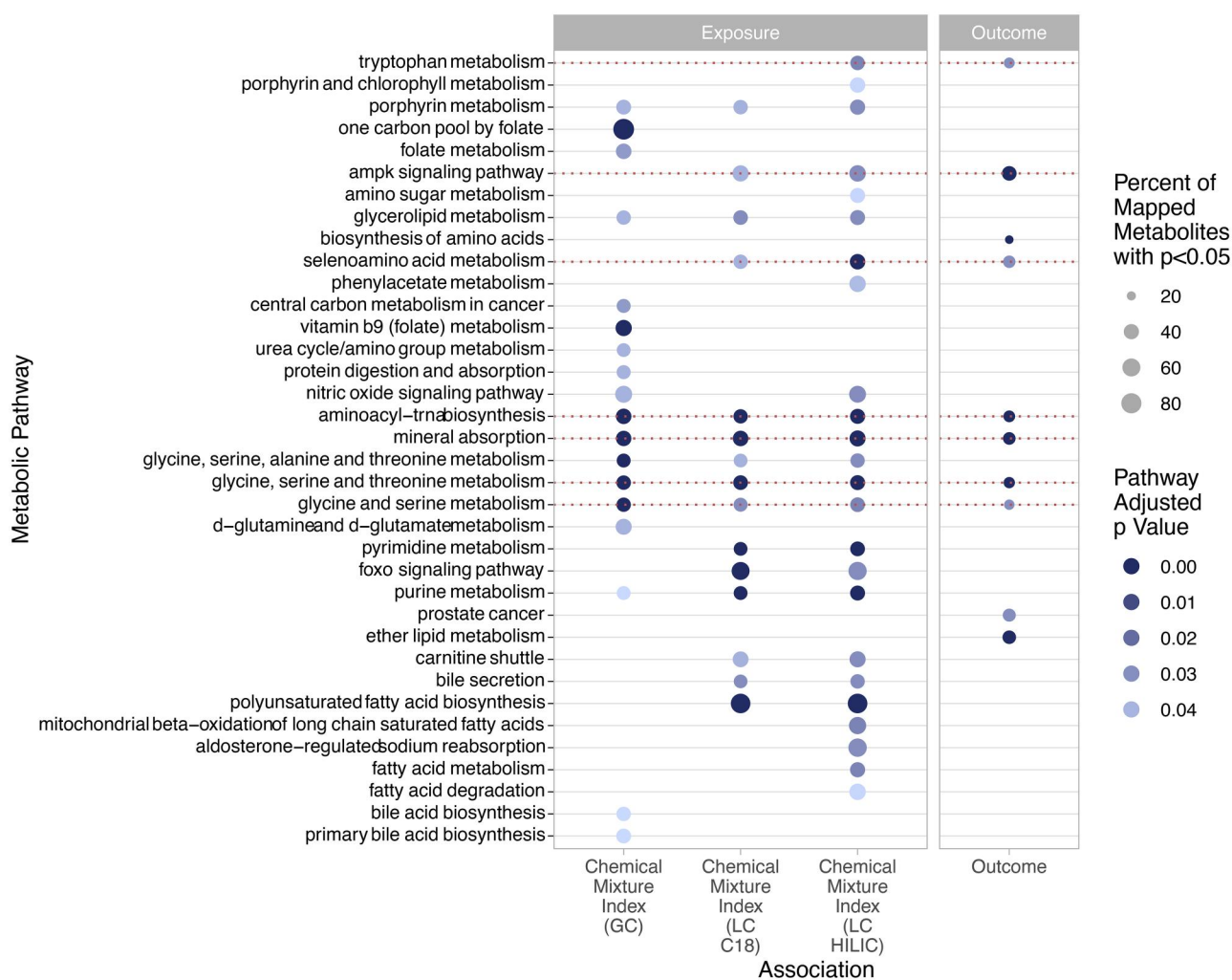
### Method decisions for metapone

When using *Metapone*, a high number of permutations, such as 1,000, will help stabilize results. The list of adducts ions to consider in annotations can be determined in consultation with the laboratory producing the data. The R package comes with a default data frame of pathway information, however, a data frame with a flag to filter to only the human pathways (flag=1) is

available at [github.com/EMERGE-EXPOSOME/Metapone](https://github.com/EMERGE-EXPOSOME/Metapone). The default threshold for significance of features within a pathway is raw  $p < 0.05$ , and the overall pathway *p* value has an optional adjusted value (by the Benjamini-Hochberg procedure) to account for multiple testing and reduce false positives. We recommend that significant pathways are also filtered to those with at least three significant metabolites (fractional counts) and that thus more comprehensively represent biological activity. The interpretations about specific pathways should be appropriately caveated due to the exploratory nature of this type of analysis. Finally, when conducting pathway enrichment using the WQSRs mixture index, it is possible to determine significantly enriched pathways *within* each repeated holdout of WQSRs (instead of using the average index) and investigate how frequently each pathway is significant; however, this would be computationally intensive.

### Additional approaches

While the MITM pathway enrichment analysis with WQS mixtures can reveal relevant mechanisms overall, the summary mixture indices may mask some significant biological pathways due to the diverse modes of action through which different chemicals operate. For this reason, it may also be useful to investigate the potential metabolic role of specific individual chemicals. For example, a few of the top contributors to the WQS mixture effect could be selected for additional pathway enrichment on their own. As another example, network analyses can explore correlations between the important chemicals in the WQS mixture effect (such as the “possible” and “probable” contributors) and each significant disease-associated metabolic pathway, where a pathway is represented by the first component (PC1) in principal component analysis (PCA) of the metabolites in that pathway ([Figure 6](#)). The PCA could be performed on only the significant



**Figure 5.** Illustrative example of how to display results of meet-in-the-middle pathway enrichment, where pathways indicated by the dotted red line are significantly enriched for both an exposure mixture and the health outcome. These results represent four different pathway enrichment analyses conducted separately (one for each of the three WQS chemical mixture indices and the outcome). GC = gas chromatography; LC = liquid chromatography; C18 =  $C_{18}$  reverse phase chromatography; HILIC = hydrophilic interaction chromatography; WQS = weighted quantile sum regression.

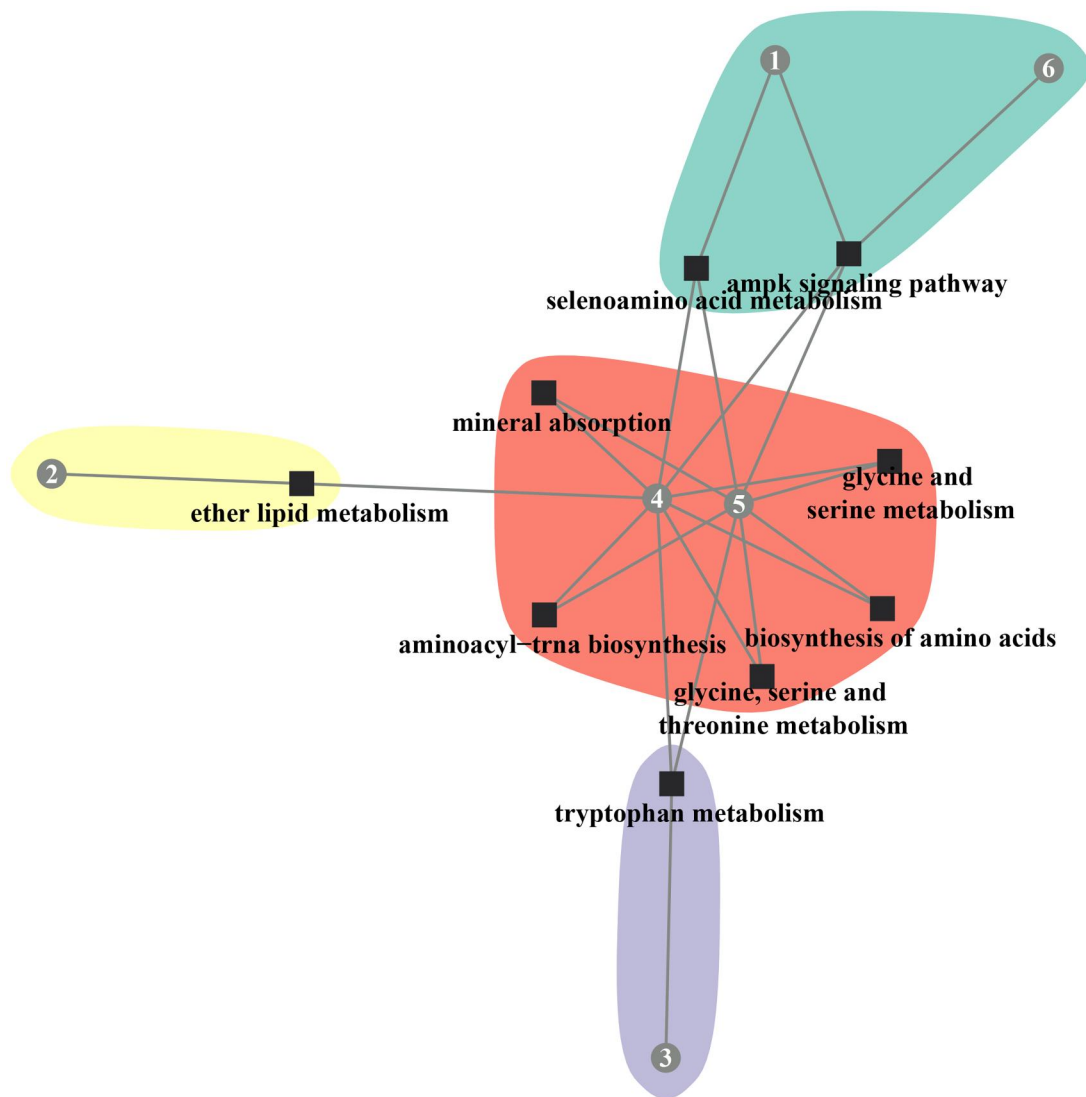
metabolites within the pathway or, to understand how chemicals impact the pathway as a whole, on all the mapped metabolites regardless of significance. It is also possible to include in the network multiple PCs per pathway (such as the subset of PCs that explain most of the variance), or to use another method for creating a summary index of the pathway. It is important to note that correlation networks are exploratory and not suitable for causal interpretation, especially given the multicollinearity issues between chemicals, so further research would be required to confirm mechanisms of action of chemicals.<sup>67</sup>

High-dimensional mediation analyses offer an alternative approach to investigate effects of exposure on disease through multiple metabolites (or preferably, pathway groups) and may improve causal inference depending on assumptions met.<sup>67-70</sup> Again, the WQS index could be helpful here as a single exposure variable representing the mixture effect, while the metabolome is retained as high-dimensional. The fact that WQS is a supervised approach that takes into account the outcome should be acknowledged when using the index to represent exposure. Other high-dimensional mediation research has employed separate exposure risk scores for different classes of chemicals,<sup>68</sup> however, this would be challenging with untargeted data where

not all chemicals are identifiable or groupable. Pairwise mediation analyses between each chemical and each metabolite or pathway group are also possible to understand specific toxicant-mediator relationships, but they are again prone to bias from chemical multicollinearity unlike mixture effect indices.<sup>68</sup>

### Strengths and limitations of pathway enrichment

Pathway enrichment is a practical exploratory approach to predict functional biological activity by leveraging pathway knowledge and bypassing the bottleneck of metabolite identification.<sup>62,64</sup> It produces interpretable and parsimonious results by mapping individual metabolites into pathway groups, which can generate hypotheses for future experimental research with the ultimate goal of identifying chemical mechanisms, therapeutic targets, or early biomarkers of disease. There are several limitations to note. Pathways are not mutually exclusive, and many metabolites are involved in multiple pathways. Thus, effects seen with a given metabolite do not always mean effects with all its pathways. The reliance only on  $m/z$  (and retention time in some cases) also limits the accuracy of metabolite annotations and may lead to false discovery, although the network mapping does help filter out the randomly structured and thus potentially



**Figure 6.** Illustrative example of a network analysis of significant, strong correlations between the disease-associated metabolic pathways (represented by their first principal component) and the exposures to chemicals that were deemed important in the weighted quantile sum (WQS) mixture effect on disease. Numbered circles refer to the chemicals and black squares to the pathways. Network was created with the R package *igraph* and clusters were determined based on multilevel community detection. Non-significant or weak correlations were not retained in the network.

irrelevant matches. In addition, the significance of enriched pathways does not reveal whether the involvement was harmful or beneficial to the health outcome, which is a challenge because of the different directions in which metabolites of the same pathway may act. Results are sensitive to the choice of databases, some of which overlook xenobiotic pathways,<sup>71</sup> and to the selected thresholds for the significance and size of pathways.<sup>72</sup> Furthermore, any pathway enrichment relies on known pathway definitions, which are subjective in their method of imposing order onto a biochemical network.<sup>72</sup> Finally, if MITM is conducted on cross-sectional exposome-metabolome data, there is the possibility for reverse causation, which limits interpretation. In general, it is best to treat pathway enrichment results as exploratory.

## Conclusion

Statistically analyzing chemicals as mixtures is important not only to capture the real-world accumulation of health burden from simultaneous exposures, but also to minimize bias from

chemical multicollinearity and co-exposure confounding. Many mixture methods address this challenge, but few currently scale to high-dimensional untargeted chemical exposome data wherein the number of features is much higher than the number of samples. WQS regression with the random subsets implementation is a statistically powerful mixture method that evaluates cumulative mixture effects on a health outcome and reveals important individual chemical drivers of the mixture effects, without loss of data resolution or interpretability. Its repetitions to estimate the mixture index across many random smaller subsets of chemicals serve to de-correlate even high-dimensional exposure data while avoiding bias and overfitting. This represents a critical advancement in the exposomics field's ability to investigate cumulative health risk from very large mixtures of exposures and to uncover emerging environmental risk factors of concern, including untargeted chemicals that are not commonly measured or not yet identifiable.

Furthermore, the cumulative mixture index can be used as a single variable representing the weighted sum of outcome-relevant exposures in integrations with other high-dimensional

omics, such as meet-in-the-middle metabolomic pathway enrichment analysis or mediation analysis. These exploratory multi-omics approaches can reveal insights into potential underlying modes of action of the chemical exposures in association with the health outcome and thus generate new hypotheses for future mechanistic research. Many decisions are required for WQS<sub>RS</sub> and pathway enrichment, so we suggest careful consideration of the discussed method parameters and customizations and recommend implementing sensitivity analyses to ensure that any conclusions are not overly sensitive to the decision points. In addition, multiple different methods may be used to test whether certain assumptions were met (such as the presence of interactions between exposures). Interpretations should also be appropriately caveated depending on the level of temporal causality in the study design. Finally, the field has continuous advancements in statistical methods, so we recommend staying attuned to new updates and functionalities for high-dimensional exposome data. In conclusion, with the ability of untargeted HRMS to now detect over 100,000 chemical signals in human samples, novel data science approaches such as WQS<sub>RS</sub> that embrace the full dimensions of the data are critical to support discovery-based exposome epidemiology and multi-omics integration.

## Author contributions

Anna S. Young (Conceptualization [equal], Methodology [equal], Funding acquisition [equal], Software [equal], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Chris Gennings (Conceptualization [equal], Methodology [equal], Resources [equal], Software [equal], Writing—review & editing [equal]), Donghai Liang (Methodology [equal], Writing—review & editing [equal]), Stephanie M. Eick (Methodology [equal], Writing—review & editing [equal]), Douglas I. Walker (Conceptualization [equal], Funding acquisition [lead], Methodology [equal], Resources [equal], Supervision [lead], Writing—review & editing [equal])

## Funding

This work was supported by the National Institute of Environmental Health Sciences at the National Institutes of Health (R01ES032831 to D.I.W. and A.S.Y., K99ES036289 to A.S.Y., K01ES035082 to S.M.E., R01ES035738 to D.L. and S.M.E., U2CES026555 to C.G., P30ES023515 to C.G.) and the National Institute of General Medical Sciences at the National Institutes of Health (R25GM143298 to A.S.Y.). The funders did not play a role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Diagrams in this paper were created using Biorender.com.

## Conflicts of interest

The authors declare no conflicts of interest.

## Data availability

No new data were generated or analyzed in support of this research.

## References

1. Wang ZY, Walker GW, Muir DCG, Nagatani-Yoshida K. Toward a global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories. *Environ Sci Technol*. 2020;54:2575-2584. <https://doi.org/10.1021/acs.est.9b06379>
2. Wagner M, Monclús L, Arp HPH, et al. State of the Science on Plastic Chemicals—Identifying and Addressing Chemicals and Polymers of Concern. Zenodo; 2024. <https://doi.org/10.5281/zenodo.10701706>
3. Wiesinger H, Wang Z, Hellweg S. Deep dive into plastic monomers, additives, and processing aids. *Environ Sci Technol*. 2021;55:9339-9351. <https://doi.org/10.1021/acs.est.1c00976>
4. Zimmerman JB, Anastas PT. Toward substitution with no regrets. *Science*. 2015;347:1198-1199. <https://doi.org/10.1126/science.aaa0812>
5. Zota AR, Calafat AM, Woodruff TJ. Temporal trends in phthalate exposures: findings from the National Health and Nutrition Examination Survey, 1–2010. *Environ Health Perspect*. 2014;122:235-241. <https://doi.org/10.1289/ehp.1306681>
6. Birnbaum LS, Bergman Å. Brominated and chlorinated flame retardants: the San Antonio statement. *Environ Health Perspect*. 2010;118:A514-A515. <https://doi.org/10.1289/ehp.1003088>
7. Brase RA, Mullin EJ, Spink DC. Legacy and emerging per- and polyfluoroalkyl substances: analytical techniques, environmental fate, and health effects. *Int J Mol Sci*. 2021;22:995. <https://doi.org/10.3390/ijms22030995>
8. US EPA. PFAS: V2 PFAS Master List of PFAS Substances. 2020. <https://comptox.epa.gov/dashboard/chemical-lists/PFASMASTERLISTV2>
9. Krahl PL, Benchoff E, Go YM, et al. Advances in comprehensive exposure assessment: opportunities for the US military. *J Occup Environ Med*. 2019;61:S5-S14. <https://doi.org/10.1097/JOM.0000000000001677>
10. Zhang P, Carlsten C, Chaleckis R, et al. Defining the scope of exposome studies and research needs from a multidisciplinary perspective. *Environ Sci Technol Lett*. 2021;8:839-852. <https://doi.org/10.1021/acs.estlett.1c00648>
11. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology, Biomarkers & Prevention* 2005;14:1847-1850. <https://doi.org/10.1158/5-9965.EPI-5-0456>
12. Miller GW. Exposomics: perfection not required. *Exposome* 2024; 4:osae006. <https://doi.org/10.1093/exposome/osae006>
13. Balcells C, Xu Y, Gil-Solsona R, Maitre L, Gago-Ferrero P, Keun HC. Blurred lines: crossing the boundaries between the chemical exposome and the metabolome. *Curr Opin Chem Biol*. 2024;78:102407. <https://doi.org/10.1016/j.cbpa.2023.102407>
14. David A, Chaker J, Price EJ, et al. Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ Int*. 2021;156:106630. <https://doi.org/10.1016/j.envint.2021.106630>
15. Walker DI, Valvi D, Rothman N, Lan Q, Miller GW, Jones DP. The metabolome: a key measure for exposome research in epidemiology. *Curr Epidemiol Rep*. 2019;6:93-103.
16. Jones DP, Cohn BA. A vision for exposome epidemiology: The pregnancy exposome in relation to breast cancer in the Child Health and Development Studies. *Reprod Toxicol*. 2020;92:4-10. <https://doi.org/10.1016/j.reprotox.2020.03.006>
17. Uppal K, Walker DI, Liu K, Li S, Go YM, Jones DP. Computational metabolomics: a framework for the million metabolome. *Chem*

- Res Toxicol. 2016;29:1956-1975. <https://doi.org/10.1021/acs.chemrestox.6b00179>
18. Liu KH, Nellis M, Uppal K, et al. Reference standardization for quantification and harmonization of large-scale metabolomics. *Anal Chem*. 2020;92:8836-8844. <https://doi.org/10.1021/acs.analchem.0c00338>
  19. Chen YC, Hsu JF, Chang CW, et al. Connecting chemical exposome to human health using high-resolution mass spectrometry-based biomonitoring: Recent advances and future perspectives. *Mass Spectrom Rev*. 2023;42:2466-2486. <https://doi.org/10.1002/mas.21805>
  20. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: where chemistry meets biology. *Science*. 2020;367:392-396. <https://doi.org/10.1126/science.aay3164>
  21. Young AS, Gennings C, Braselton ME, et al. Integrated chemical exposome-metabolome profiling of follicular fluid and associations with fertility outcomes during assisted reproduction. *Environ Int*. 2025;203:109787. <https://doi.org/10.1016/j.envint.2025.109787>
  22. Joubert BR, Kioumourtoglou MA, Chamberlain T, et al. Powering research through innovative methods for mixtures in epidemiology (prime) program: novel and expanded statistical methods. *Int J Environ Res Public Health*. 2022;19. <https://doi.org/10.3390/ijerph19031378>
  23. Kienzler A, Bopp SK, van der Linden S, Berggren E, Worth A. Regulatory assessment of chemical mixtures: Requirements, current approaches and future perspectives. *Regulatory. Regul Toxicol Pharmacol*. 2016;80:321-334. <https://doi.org/10.1016/j.yrtph.2016.05.020>
  24. Kortenkamp A. Low dose mixture effects of endocrine disruptors and their implications for regulatory thresholds in chemical risk assessment. *Curr Opin Pharmacol*. 2014;19:105-111. <https://doi.org/10.1016/j.coph.2014.08.006>
  25. Patel CJ. Analytic complexity and challenges in identifying mixtures of exposures associated with phenotypes in the exposome era. *Curr Epidemiol Rep*. 2017;4:22-30. <https://doi.org/10.1007/s40471-7-0100-5>
  26. Weisskopf MG, Seals RM, Webster TF. Bias amplification in epidemiologic analysis of exposure to mixtures. *Environ Health Perspect*. 2018;126:047003. <https://doi.org/10.1289/EHP2450>
  27. Stapleton HM, Klosterhaus S, Eagle S, et al. Detection of organophosphate flame retardants in furniture foam and U.S. house dust. *Environ Sci Technol*. 2009;43:7490-7495. <https://doi.org/10.1021/es9014019>
  28. Levin R, Villanueva CM, Beene D, et al. US drinking water quality: exposure risk profiles for seven legacy and emerging contaminants. *J Expo Sci Environ Epidemiol*. 2024;34:3-22. <https://doi.org/10.1038/s41370-3-00597-z>
  29. Johns LE, Cooper GS, Galizia A, Meeker JD. Exposure assessment issues in epidemiology studies of phthalates. *Environ Int*. 2015; 85:27-39. <https://doi.org/10.1016/j.envint.2015.08.005>
  30. Tu YK, Gunnell D, Gilthorpe MS. Simpson's paradox, lord's paradox, and suppression effects are the same phenomenon—the reversal paradox. *Emerg Themes Epidemiol* 2008;5:2. <https://doi.org/10.1186/2-7622-5-2>
  31. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J Agric Biol Environ Stat*. 2015;20:100-120. <https://doi.org/10.1007/s13253-4-0180-3>
  32. Braun JM, Gennings C, Hauser R, Webster TF. What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environ Health Perspect*. 2016;124:A6-A9. <https://doi.org/10.1289/ehp.1510569>
  33. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 2013;36:27-46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
  34. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale)*. 2016;6:227. <https://doi.org/10.4172/2161-1165.1000227>
  35. Chung MK, House JS, Akhtari FS, Members of the Exposomics Consortium, et al. Decoding the exposome: data science methodologies and implications in exposome-wide association studies (ExWASs). *Exposome* 2024;4:osae001. <https://doi.org/10.1093/exposome/osae001>
  36. Li S, Cirillo P, Hu X, et al. Understanding mixed environmental exposures using metabolomics via a hierarchical community network model in a cohort of California women in 1960's. *Reprod Toxicol*. 2020;92:57-65. <https://doi.org/10.1016/j.reprotox.2019.06.013>
  37. Zhu G, Wen Y, Cao K, He S, Wang T. A review of common statistical methods for dealing with multiple pollutant mixtures and multiple exposures. *Front Public Health*. 2024;12:1377685. <https://doi.org/10.3389/fpubh.2024.1377685>
  38. Pan S, Li Z, Rubbo B, et al. Applications of mixture methods in epidemiological studies investigating the health impact of persistent organic pollutants exposures: a scoping review. *J Expo Sci Environ Epidemiol*. 2025;35:522-534. Published online September 10. <https://doi.org/10.1038/s41370-4-00717-3>
  39. Bobb JF, Valeri L, Claus Henn B, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*. 2014;16:493-508. <https://doi.org/10.1093/biostatistics/kxu058>
  40. Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ Health*. 2018;17: 67. <https://doi.org/10.1186/s12940-8-0413-y>
  41. Gibson EA, Nunez Y, Abuawad A, et al. An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length. *Environ Health*. 2019;18:76. <https://doi.org/10.1186/s12940-9-0515-1>
  42. Joubert BR, Palmer G, Dunson D, Kioumourtoglou MA, Coull BA. Workflow for statistical analysis of environmental mixtures. *Environ Health Perspect*. 2025. <https://doi.org/10.1289/EHP16791>
  43. Keil AP, Buckley JP, O'Brien KM, Ferguson KK, Zhao S, White AJ. A quantile-based g-computation approach to addressing the effects of exposure mixtures. *Environ Health Perspect*. 2020;128: 47004. <https://doi.org/10.1289/EHP5838>
  44. Gennings C. Comment on "a quantile-based g-computation approach to addressing the effects of exposure mixtures". *Environ Health Perspect*. 2021;129:38001. <https://doi.org/10.1289/EHP8739>
  45. Renzetti S, Gennings C, Calza S. A weighted quantile sum regression with penalized weights and two indices. *Front Public Health*. 2023;11:1151821. Accessed August 21, 2023. <https://www.frontiersin.org/articles/10.3389/fpubh.2023.1151821>
  46. Tanner EM, Bornehag CG, Gennings C. Repeated holdout validation for weighted quantile sum regression. *MethodsX* 2019;6: 2855-2860. <https://doi.org/10.1016/j.mex.2019.11.008>
  47. Hao W, Cathey AL, Aung MM, Boss J, Meeker JD, Mukherjee B. Statistical methods for chemical mixtures: a roadmap for practitioners using simulation studies and a sample data analysis in the PROTECT cohort. *Environ Health Perspect*. 2025;133:67019. Published online May 20. <https://doi.org/10.1289/EHP15305>

48. Kalia V, Walker DI, Krasnodemski KM, Jones DP, Miller GW, Kioumourtzoglou MA. Unsupervised dimensionality reduction for exposome research. *Curr Opin Environ Sci Health*. 2020;15:32-38. <https://doi.org/10.1016/j.coesh.2020.05.001>
49. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970;12:55-67. <https://doi.org/10.1080/00401706.1970.10488634>
50. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2005;67:301-320. <https://doi.org/10.1111/j.7-9868.2005.00503.x>
51. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58:267-288. <https://doi.org/10.1111/j.7-6161.1996.tb02080.x>
52. Czarnota J, Gennings C, Colt JS, et al. Analysis of environmental chemical mixtures and non-Hodgkin lymphoma risk in the NCI-SEER NHL study. *Environ Health Perspect*. 2015;123:965-970. <https://doi.org/10.1289/ehp.1408630>
53. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat*. 2009;37:1733-1751. <https://doi.org/10.1214/08-AOS625>
54. Curtin P, Kellogg J, Cech N, Gennings C. A random subset implementation of weighted quantile sum (WQSRS) regression for analysis of high-dimensional mixtures. *Communications in Statistics—Simulation and Computation*. 2021;50:1119-1134. <https://doi.org/10.1080/03610918.2019.1577971>
55. Mohammed Taha H, Aalizadeh R, Alygizakis N, et al. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. *Environ Sci Eur*. 2022;34:104. <https://doi.org/10.1186/s12302-2-00680-6>
56. Uppal K, Walker DI, Jones DP. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. *Anal Chem*. 2017;89:1063-1067. <https://doi.org/10.1021/acs.analchem.6b01214>
57. Bennett DH, Busgang SA, Kannan K, et al. Environmental exposures to pesticides, phthalates, phenols and trace elements are associated with neurodevelopment in the CHARGE study. *Environ Int*. 2022;161:107075. <https://doi.org/10.1016/j.envint.2021.107075>
58. Busgang SA, Spear EA, Andra SS, et al. Application of growth modeling to assess the impact of hospital-based phthalate exposure on preterm infant growth parameters during the neonatal intensive care unit hospitalization. *Sci Total Environ*. 2022;850:157830. <https://doi.org/10.1016/j.scitotenv.2022.157830>
59. Campbell RK, Curtin P, Enlow MB, Brunst KJ, Wright RO, Wright RJ. Disentangling associations among maternal lifetime and prenatal stress, psychological functioning during pregnancy, maternal race/ethnicity, and infant negative affectivity at age 6 months: a mixtures approach. *Health Equity*. 2020;4:489-499. <https://doi.org/10.1089/heap.2020.0032>
60. Invernizzi A, Rechtman E, Curtin P, et al. Functional changes in neural mechanisms underlying post-traumatic stress disorder in World Trade Center responders. *Transl Psychiatry*. 2023;13:239-10. <https://doi.org/10.1038/s41398-3-02526-y>
61. Babin É, Cano-Sancho G, Vigneau E, Antignac JP. A review of statistical strategies to integrate biomarkers of chemical exposure with biomarkers of effect applied in omic-scale environmental epidemiology. *Environ Pollut*. 2023;330:121741. <https://doi.org/10.1016/j.envpol.2023.121741>
62. Li S, Park Y, Duraisingham S, et al. Predicting network activity from high throughput metabolomics. *PLOS Comput Biol*. 2013;9:e1003123. <https://doi.org/10.1371/journal.pcbi.1003123>
63. Pang Z, Chong J, Zhou G, et al. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res*. 2021;49:W388-W396. <https://doi.org/10.1093/nar/gkab382>
64. Tian L, Li Z, Ma G, et al. Metapone: a Bioconductor package for joint pathway testing for untargeted metabolomics data. *Bioinformatics*. 2022;38:3662-3664. <https://doi.org/10.1093/bioinformatics/btac364>
65. Ogata H, Goto S, Fujibuchi W, Kanehisa M. Computation with the KEGG pathway database. *Biosystems*. 1998;47:119-128. [https://doi.org/10.1016/S0303-2647\(98\)00017-3](https://doi.org/10.1016/S0303-2647(98)00017-3)
66. Frolkis A, Knox C, Lim E, et al. SMPDB: the small molecule pathway database. *Nucleic Acids Res*. 2010;38:D480-D487. <https://doi.org/10.1093/nar/gkp1002>
67. Fuller H, Zhu Y, Nicholas J, et al. Metabolomic epidemiology offers insights into disease aetiology. *Nat Metab*. 2023;5:1656-1672. <https://doi.org/10.1038/s42255-3-00903-x>
68. Aung MT, Song Y, Ferguson KK, et al. Application of an analytical framework for multivariate mediation analysis of environmental data. *Nat Commun*. 2020;11:5624. <https://doi.org/10.1038/s41467-0-19335-2>
69. Goodrich JA, Wang H, Jia Q, et al. Integrating Multi-Omics with environmental data for precision health: A novel analytic framework and case study on prenatal mercury induced childhood fatty liver disease. *Environ Int*. 2024;190:108930. <https://doi.org/10.1016/j.envint.2024.108930>
70. Zhang H, Zheng Y, Zhang Z, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*. 2016;32:3150-3154. <https://doi.org/10.1093/bioinformatics/btw351>
71. Liang D, Li Z, Vlaanderen J, et al. A state-of-the-science review on high-resolution metabolomics application in air pollution health research: current progress, analytical challenges, and recommendations for future direction. *Environ Health Perspect*. 2023;131:56002. <https://doi.org/10.1289/EHP11851>
72. Wieder C, Bundy JG, Frainay C, et al. Avoiding the misuse of pathway analysis tools in environmental metabolomics. *Environ Sci Technol*. 2022;56:14219-14222. <https://doi.org/10.1021/acs.est.2c05588>

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Exposome, 2025, 5, 1–14

<https://doi.org/10.1093/exposome/osaf010>

Commentary