Statistics in Biosciences
https://doi.org/10.1007/s12561-023-09405-6

Detecting Shape-Based Interactions Among Environmental Chemicals Using an Ensemble of Exposure-Mixture Regression and Interpretable Machine Learning Tools

Vishal Midya[1] · Chris Gennings[1]

Received: 29 May 2023 / Revised: 2 September 2023 / Accepted: 12 October 2023

Abstract

There is growing interest in discovering interactions between multiple environmental chemicals associated with increased adverse health effects. However, most existing approaches (1) either use a projection or product of multiple chemical exposures, which are difficult to interpret and (2) cannot simultaneously handle multi-ordered interactions. Therefore, we develop and validate a method to discover shape-based interactions that mimic usual toxicological interactions. We developed the Multi-ordered explanatory interaction (Moxie) algorithm by merging the efficacy of Extreme Gradient Boosting with the inferential power of Weighted Quantile Sum regression to extract synergistic interactions associated with the outcome/odds of disease in an adverse direction. We evaluated the algorithm's performance through simulations and compared it with the currently available gold standard, the signed-iterative random forest algorithm. We used the 2017–18 US-NHANES dataset (n = 447 adults) to evaluate interactions among nine per- and poly-fluoroalkyl substances and five metals measured in whole blood in association with serum low-density lipoprotein cholesterol. In simulations, the Moxie algorithm was highly specific and sensitive and had very low false discovery rates in detecting true synergistic interactions of 2nd, 3rd, and 4th order through moderate (n = 250) to large (n = 1000) sample sizes. In NHANES data, we found a two-order synergistic interaction between cadmium and lead detected in people with whole-blood cadmium concentrations and lead above 0.605 ug/dL and 1.485 ug/dL, respectively. Our findings demonstrate a novel validated approach in environmental epidemiology for detecting shape-based toxicologically mimicking interactions by integrating exposure-mixture regression and machine learning methods.

Keywords Exposure mixture models · Weighted quantile sum regression · Interactions · Machine learning · Extreme gradient boosting

Extended author information available on the last page of the article

Published online: 23 November 2023

1 Introduction

There is a growing interest in evaluating the impact of mixtures of environmental chemicals and the interactions that might occur among their components, particularly with human disease endpoints. Humans are simultaneously exposed to multiple chemicals; therefore, chemical exposures that share similar stereo-chemical properties, or have steady bioaccumulation, or lie in similar biological pathways may lead to interactions [1] and [2]. Nevertheless, from biological sciences to chemistry, epidemiology to mathematical statistics, the notion of "interaction" has been interpreted through different paradigms [3, 4]. In current environmental epidemiology and biostatistics literature, interactions are usually presented as a projection (for example, product) of multiple exposures. To evaluate interactions, some models "hard-code" or pre-specify interaction terms [5–8]. In contrast, another group of models uses flexible non-parametric or semi-parametric tools to estimate the exposure-response surface and the overall mixture effect and, therefore, could model interactions as a byproduct of their model [9–11]. However, their qualitative graphical assessment makes it difficult to reach any definite conclusion, particularly when the order of the interactions or the number of exposures is large [12]. In the end, these projected terms are therefore treated as interactions, and their effect sizes (or inclusion probabilities) are estimated (usually in the presence of the main exposures under strong or heredity assumptions) [13]. Although this perspective had led to sophisticated and flexible computational techniques, these tools were not primed to shed light on mechanistic or toxicological insights into how the interactions could lead to biochemical plausibility.

On the contrary, in pharmacology/toxicology literature, more priority has been given to elucidating interactions with direct biochemical implications. For example, among several others, the idea of "dose addition" [4, 14–17] was used to interpret interactions and quantify departure from zero interactions. The concept of "dose addition" implied that when two chemicals produced similar effects, then the joint effect of their combination might either be equal to the algebraic sum of their effect (additive) or might be more (synergistic) or less (antagonistic) than the algebraic sum [18]. Such interactions primarily focus on bio-chemical plausibility, mechanistic insights, and the directionality of associations. Tree-based Machine-learning models provide a natural way to discover such interactions. Nevertheless, a major challenge is that most of the inner workings of these tree-based models are not easily interpretable, creating a tension between prediction quality and meaningful biological insight. Moreover, a predictive machine-learning model might not be the optimal model for inference [19]. In recent epidemiological studies, tree-based machine-learning tools were used to discover combinations of interacting environmental chemicals in the absence of main mixture effects, therefore possibly incurring potential ramifications in terms of false positives [20–23].

In this paper, we aimed to find synergistic interactions that are adversely associated with the outcome in the presence of a chemical mixture index. We developed an algorithm to search for synergistic interactions without investigating each possible combination. Using a blend of Weighted Quantile Sum Regression [5] and Extreme gradient boosting [24], this algorithm develops certain heuristic rules to discover toxicologically mimicking interactions that provide interpretability and decrease the computational burden. We named this algorithm as (M) ulti-(o)rdered e(x)planatory (I)nt(e)ractions or "Moxie" algorithm. We established certain properties of the Moxie algorithm and conducted extensive simulations to compare and contrast its properties with the current gold standard. Finally, we used this algorithm to discover interactions between Perfluoroalkyl substances and metal exposures associated with serum low-density lipoprotein concentrations among adults from the 2017-18 US National Health and Nutrition Examination Survey (NHANES).

2 Modeling Approach

We developed the multi-ordered explanatory Interaction (Moxie) algorithm, which combined the Weighted Quantile Sum (WQS) regression [5], Extreme Gradient Boosting (XGBoost) [24], along with a few layers of novel heuristic techniques to discover synergistic chemical interactions in association with disease endpoints. First, this algorithm leveraged the WQS regression to construct a chemical mixture index. Second, on top of the main chemical-mixture index, the algorithm utilized XGBoost to extract all potential synergistic interactions and carefully employed a few heuristic rules to sieve the true signals. Below, we describe the stages in more detail.

2.1 Weighted Quantile Sum Regression: Creating Exposure-Mixture Index

We started by fitting a generalized WQS model [5] with a chosen set of chemical exposures and potential covariates to model the simultaneous mixture of exposure to environmental chemicals. We focused on WQS as a mixture model because of the prior hypothesis on the directionality of association, its robustness in implementation, and simplicity in interpreting the mixture index. Several other exposure-mixture methods [6, 7, 9] could also be used to construct the mixture index, but the aims and interpretations will vary. In this current algorithm, the WQS model was implemented with a random subset [25] and repeated holdouts validation [26]. The mixture index obtained from this model was treated as the main effect. After controlling for the mixture index and covariates, we extracted the residuals from this model, assuming that this residual included potentially informative synergistic effects. The following stages aimed to extract synergistic interactions associated with the outcome in an adverse direction based on the residuals. A detailed discussion on WQS-type models can be found in Joubert et al. [27].

2.2 Extreme Gradient Boosting: Constructing Shallow Trees with Potential Synergistic Interactions

XGBoost [24] is a particular gradient-boosted algorithm that uses an ensemble technique to grow shallow and iterative decision trees—a key difference from a random forest [28]. Further, on top of the usual gradient-boosted trees, XGBoost included penalization and state-of-the-art data engineering techniques for rapidly generating trees. On the residual from WQS, we fitted XGBoost models to grow a forest of shallow, iteratively learned decision trees. Since explanation, not prediction, was the aim, we chose not to focus on prediction quality; instead, we used random subsets of the whole dataset to fit many models and used all generated trees, irrespective of their predictive power [19].

Consider a data frame with n participants, p chemical exposures, and y be the outcome. Let $\mathbb{D} = \{(\boldsymbol{x_i}, y_i)\}$, with $|\mathbb{D}| = n$, $\boldsymbol{x_i} \in \mathbb{R}^p$, and, $y_i \in \mathbb{R}$. Then, a tree ensemble model in XGBoost takes the form

$$\hat{y}_i = \phi(\boldsymbol{x_i}) = \sum_{k=1}^{K} f_k(\boldsymbol{x_i}), f_k \in \mathbb{F},$$

where K denotes the number of additive functions to predict the y_i, which is the residual from subsection 2.1. Further, (1) $\mathbb{F} = \{f(x) = w_{q(x)}\}(q : \mathbb{R}^p \rightarrow \{1, 2, \ldots, T\}, w \in \mathbb{R}^T)$ $\mathbb{F} = \{f(x) = w_{q(x)}\}(q : \mathbb{R}^p \rightarrow \{1, 2, \ldots, T\}, w \in \mathbb{R}^T)$, (2) q denotes the structure of each tree that maps an individual sample to a corresponding leaf index, (3) T denotes the number of leaves in the tree, and (4) f_k denotes independent tree structure q and leaf weight w. Borrowing the formulation from Gelfand et al. [29], a tree, f_k is built upon a set of nodes, t, and two functions left- I_L and right-I_R as $I_L, I_R : f_k \rightarrow f_k \bigcup \{0\}$ such that

1. $\forall t \in f_k$, either $I_L(t) = 0$ and $I_R(t) = 0$ or $I_L(t) > t$ and $I_R(t) > t$.
2. $\forall t \in f_k$, other than the smallest integer in f_k, there is an unique $s \in f_k$, such that either $t = I_L(s)$ or $t = I_R(s)$.
3. Moreover, without loss of generalizability, $I_L(t) = I_R(t)$ when $I_L(t) = 0$ and denoted by $root(f_k)$. A node will be called a terminal node if $I_L(t) = 0$, else it is a non-terminal node, and if $I_R(t) = I_L(t) + 1$ when $I_L(t) > 0$, then the tree can be denoted solely by I_L.

In the Moxie algorithm, we aim to grow K additive and shallow trees, repetitively created over $m(m = 1(1)M)$ repetitions of bootstrapped samplings, and at each of the tree-depths of $d(d = 1(1)D)$. We denote the k^{th} tree grown at m^{th} iteration by the notation $f_{k,m,\cup d}$, with $\cup d$ denoting the collection of multiple tree depths, $d = 1, \ldots, D$. The h^{th} node in the tree $f_{k,m,\cup d}$ is noted as $t^h_{f_{k,m,\cup d}}$. Then for root node, $I_R(t^0_{f_{k,m,\cup d}}) = I_L(t^0_{f_{k,m,\cup d}})$ for $root(f_{k,m,\cup d})$. A branch of length H in a tree $f_{k,m,\cup d}$ is defined as an ordered set of nodes $\left\{ t^h_{f_{k,m,\cup(d)}} \right\}_{h=0}^{H}$ connected through a sequence of integers from the parent to a child node, such that there is a maximum of one node per branch at each depth of the tree. The following properties hold for a tree:

- $t^{0}_{f_{k,m,d=0}} = root(f_{k,m,\cup d})$.
- If $t^{h_l}_{f_{k,m,d_i}}$ and $t^{h_q}_{f_{k,m,d_j}}$ belong to the same branch, and if $h_l \neq h_q$ and $d_j = d_i + \theta$, then $t^{h_q}_{f_{k,m,d_j}} = child^{\theta}(t^{h_l}_{f_{k,m,d_i}})$, where, the operator $child^{\theta}$ denotes θ generations have passed from the parent $t^{h_l}_{f_{k,m,d_i}}$.
- And finally, $I_L(t^{h=H}_{f_{k,m,d=D}}) = 0$ and $I_R(t^{h=H}_{f_{k,m,d=D}}) = 0$.

This paper concerns synergistic interactions in an adverse direction (i.e., more than the additive effect on top of higher concentrations). Therefore, we will focus on the right branches of the grown trees. We further define the notation of right branch for a tree $f_{k,m,\cup d}$ (with length of the branch being H) as, $B_R(f_{k,m,\cup d})$ as an ordered set of nodes $\left\{\bigcup_{h=0}^{H} t^{h}_{f_{k,m,\cup(d)}}\right\}$ such that if $t^{h_l}_{f_{k,m,d_i}} \in B_R(f_{k,m,\cup d})$ and $t^{h_q}_{f_{k,m,d_j}} \in B_R(f_{k,m,\cup d})$, and $h_l \neq h_q, d_j = d_i + 1$, then $t^{h_q}_{f_{k,m,d_j}} = I_R(t^{h_l}_{f_{k,m,d_i}})$ and $t^{h_q}_{f_{k,m,d_j}} \neq I_L(t^{h_l}_{f_{k,m,d_i}})$. Moreover, a right branch is created from a node $t^{h}_{f_{k,m,d}}$ that represents an exposure and a threshold value for the split; therefore, $t^{h}_{f_{k,m,d}}$ is a set composed of the exposure and a threshold value $\{X^{h}_{f_{k,m,d}}, S^{h}_{f_{k,m,d}}\}$, where X denotes the exposure and S denotes the threshold. Intuitively, these right branches can potentially detect a more than additive interaction. We now define a function “Feature” that extracts all the exposures from a Right Branch,

$$F \bullet B_R(f_{k,m,\cup d}) := \left\{ \bigcup_{h} X^{h}_{f_{k,m,d}} \ : \ \mathbb{R}^{p} \times \mathbb{R} \to \mathbb{R}^{p} \right\}, |F \bullet B_R(f_{k,m,\cup d})| = H.$$

The “unique Feature” is a function that is denoted by $uF \bullet B_R(f_{k,m,\cup d}) := \left\{ \bigcup_{h} X^{h}_{f_{k,m,d}} \right\}_{\neq}$ $uF \bullet B_R(f_{k,m,\cup d}) := \left\{ \bigcup_{h} X^{h}_{f_{k,m,d}} \right\}_{\neq}$, selects the set of unique exposures from all the exposures $F \bullet B_R(f_{k,m,\cup d})$ within a Right Branch. Note that, $|uF \bullet B_R(f_{k,m,\cup d})| \leq |F \bullet B_R(f_{k,m,\cup d})|$. Note that there might be multiple occurrences of the same exposures in a set of features, but the corresponding thresholds will be distinct. Since we are focused on the right branches, the final threshold corresponding to the unique features will be extracted from the highest depth in a set of unique features. For example, consider A+(> a_1)/B+(> b)/C+(> c)/A+(> a_2) as a right branch from a tree grown to depth four, with the set of exposures A, B, and, C and the corresponding rules dictated by the thresholds a_1, b, c, and, a_2. Therefore, the set A+/B+/C+/A+ represents a feature, and the set A+/B+/C+ shows the unique feature.

2.3 Discriminating Behavior of Right Branches Using a Set Decomposition Technique

The set of unique features represents shape-based interactions from the right branches. However, not all unique features are equally important. First, the fitted decision trees are optimized for better prediction accuracy and, therefore, tend to grow branches that blend true-positive and false-positive features; we name these branches "pseudo features." Second, since machine learning models usually overfit, some of the right branches could be composed of completely false-positive features; we name these branches as "dead features." In the case of prediction, pseudo branches are not a curse as they aid in prediction (since they are still partially informative) but propagate false positives and false discovery rates in the case of inference. We devise and implement a few techniques described below to counteract such phenomena.

Let $\left\{\bigcup_{k,m} uF \cdot B_R(f_{k,m,\cup d})\right\}_{\neq}$ denote the set of all unique features extracted from each of K additive trees grown at each of the M iterations and $\zeta = |\left\{\bigcup_{k,m} uF \cdot B_R(f_{k,m,\cup d})\right\}_{\neq}|$. Assume, $\alpha_1, \alpha_2, \ldots, \alpha_l$ be the sets of unique features in $\left\{\bigcup_{k,m} uF \cdot B_R(f_{k,m,\cup d})\right\}_{\neq}$. Then, we define a function named "Stability' that calculates the frequency of occurrence of each unique feature.

$$Stability(\alpha_j) = \sum_{i=1}^{\zeta} I\{\alpha_j = z_i, \text{ s.t. } z_i \in \bigcup_{k,m} uF \cdot B_R(f_{k,m,\cup d})\},$$

where $I(.)$ denotes an indicator function. For all the extracted right branches, we only select those with (1) a prevalence of features that should be more than 5% (to ensure the selected features are not results from overfitted right branches), and (2) the length of each unique feature more than 1 (i.e., the right branch should have more than one exposures). Next, we convert each selected branch into indicator functions based on their concentrations (and corresponding split thresholds) to denote the presence (non-zero) or absence (zero) of interactions. For each of the indicators, we estimate corresponding beta estimates (after controlling for the WQS mixture index and covariates). Since we are interested only in synergistic interactions in an adverse direction, we restrict to those right branches whose corresponding beta estimates are positive (i.e., the beta estimates should have the same sign as the direction of adversity for synergy). Note that easing this restriction will also lead to synergistic interactions in a non-adverse direction.

2.3.1 Stability Ratios

Ideally, we would want those right branches with high stability of unique features and a high association of indicators with the outcome. Therefore, we create two sets to distinguish these right branches, as noted.

- Denote the first set as $\mathcal{S}_{B_R(f)}$. Let, $x \in \mathcal{S}_{B_R(f)}$ and $f_{k_0,m_0,d_0} \in f$, then (1) $|F \cdot x| = d_0$, and (2) $Stability(x)$ and the beta estimate of x should be higher than respective α^{th} percentiles.

- Denote the second set as $\mathcal{L}_{B_R(f)}$. Let, $x \in \mathcal{L}_{B_R(f)}$ and $f_{k_0,m_0,d_0} \in f$, then (1) $|uF \cdot x| = d_0$.

Now for each $x \in \mathcal{L}_{B_R(f)}$, we define,

$$\text{Stability Ratio}(x) = \frac{max\left\{Stability\left(uF \cdot \{\mathcal{L}_{B_R(f)} \backslash \mathcal{S}_{B_R(f)}\}\right)\right\}}{Stability(x|x \in \mathcal{L}_{B_R(f)})}.$$

A key observation is that as long as there truly is a synergistic interaction, the Stability Ratio of the corresponding right branch should be far less than one since it is calculated on the set difference $\mathcal{L}_{B_R(f)} \backslash \mathcal{S}_{B_R(f)}$. Whereas for pseudo and dead features, the Stability Ratios should be larger than one. Using two simulated scenarios, we show that when there was a true synergistic interaction, Stability Ratio < 1 induced a discrimination that could classify true signals from false ones. Figure 1 presents the distributions of Stability Ratios in two simulated scenarios for sample sizes $n = 1000$ and $n = 500$. A three-ordered synergistic interaction was embedded in the outcome, induced by exposures V_1, V_3, and V_5 on top of the main-mixture effect by V_1 and V_2. We included 50 exposures, V_1, V_2,..., V_{50}. The truly unique feature, denoted by $V_1 + /V_3 + /V_5+$, induced smaller two-ordered interactions, $V_1 + /V_3+$, $V_1 + /V_5+$, and $V_3 + /V_5+$, which we name induced true feature. Branches consisting of V_1, V_3, and/or V_5, along with any of the other exposures, for example, $V_1 + /V_3 + /V_{40}+$ or $V_3 + /V_{12}+$, are the pseudo features. Features with neither V_1, V_3, nor V_5 are called dead features, such as $V_{10} + /V_7 + /V_{40}+$. Multiple right branches have unique features $V_1 + /V_3 + /V_5+$ with separate thresholds; therefore, one can obtain a distribution of the Stability Ratios even for a single interaction.

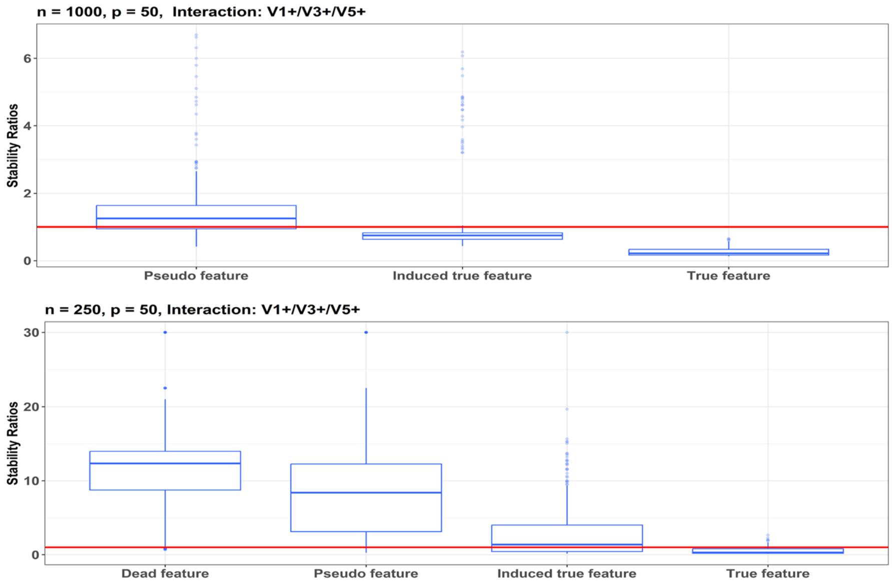

Fig. 1 Distribution of Stability Ratios for $n = 1000$ and 250 with 50 exposures

The median Stability Ratio of dead, pseudo, induced true, and true features followed a decreasing trend irrespective of the sample sizes. The median Stability Ratio for true features remained less than one, particularly as the sample size increased; the complete distribution of Stability Ratio for true features was below one. Therefore, a cutoff (< 1) on Stability Ratio could discriminate between the different kinds of features.

2.4 Exposure Co-occurrence Lists

Figure 1 shows that some dead or pseudo features could still pass the below one cutoff Stability Ratio for smaller sample sizes. Therefore, as a next step, we created a Feature co-occurrence list following the heuristics from distributed word representations (popularly known as word embedding) and widely applied in tasks related to Natural Language Processing [30] and [31]. First, we calculated the frequency of occurrence for each of the features in the selected right branches, and second, we mapped those frequencies to each of the branches to quantify the co-occurrence frequencies. For example, consider the branches $V_1+/V_3+/V_5+$, V_3+/V_5+, V_3+/V_7+, $V_5+/V_{10}+$, and, V_1+/V_3+ obtained from the previous stage. Let $V_1+/V_3+/V_5+$ be the true branch and V_3+/V_7+ and $V_5+/V_{10}+$ be the pseudo features. Note the frequencies of individual features V_1+, V_3+, V_5+, V_7+, and $V_{10}+$ to be two, four, three, one, and one, respectively. Next, we mapped these frequencies to each branch to create co-occurrence frequencies. For $V_1+/V_3+/V_5+$, V_3+/V_5+, V_3+/V_7+, $V_5+/V_{10}+$, and, V_1+/V_3+ the list of co-occurrence frequencies was 2/4/3, 4/3, 4/1, 3/1, and, 2/4. For true or induced true features, each item of the exposure co-occurrence list should ideally have frequencies of more than one, whereas, in pseudo or dead features, many of the exposures occurred just once. This was because randomly generated false-positive exposures may latch onto true or induced true features and create pseudo features since the machine-learning model was primed to optimize prediction. Therefore, we only considered those branches whose co-occurrence frequencies were at least two.

2.5 Friedman's H-statistic

Feature co-occurrence lists worked very effectively for larger samples, but in cases with smaller sample sizes and a larger number of exposures, they might still fail to sieve out pseudo or dead features. Therefore, we calculated Friedman's H-statistic [32] on the selected branches to discriminate between the true and the false ones. Friedman's H-statistic was based on a variable importance measure and quantified relative importance; see [21] for a summary. The value of the H-statistic ranged from zero to one, with larger values indicating a stronger interaction effect. Although H-statistic would not necessarily be zero in practice due to sampling fluctuations, false-positive interactions should have very small values. Even though the H-statistic was relatively easy to use, the interpretation of its interaction strength was not comparable to either effect estimates or inclusion probabilities. H-statistic could be generalized to discover interactions of any order—but it needed to be estimated for each combination by specifying the intended terms—which could be computationally intensive when the number of exposures and the order of interactions increases. Therefore, the H-statistic was very suitable to use once there were prior plausible interaction terms to test for. Note that the H-statistic does not have directional awareness and cannot distinguish between synergistic or antagonistic interactions. We only calculated the H-statistic for a few pre-selected branches in the Moxie algorithm. As the final gatekeeper, any branch with an H-statistic larger than the pre-specified cutoff was designated the final synergistic interaction.

3 Simulation for Moxie-Algorithm to Demonstrate its Ability to Detect Interactions

We conducted extensive simulations to quantify the performances and limitations of the Moxie algorithm. We compared its performance with the Signed iterative Random Forest (SiRF) [33] and [34]. The SiRF utilized a combination of iterative random forest and random intersection trees to search for informative and stable multi-order interactions [35]. Since this algorithm is based on weighted iterative random forests, the discovered multi-ordered interactions were based on thresholds and, therefore, could be interpreted to mimic toxicological interactions. Although SiRF aims for predictions, not associations, in this simulation, we compare the discovered chemical interactions by both algorithms. We restricted our comparisons to only tree-based algorithms (i.e., algorithms that provide interactions with a coarse representation of hyper-rectangles based on the decision rules and thresholds). We excluded product or projection-based algorithms that do not directly represent the collective activity of the chemical exposures. Note that improved variations of the SiRF algorithm using a repeated hold-out stage [36] and [37] have been proposed in the literature. But we simply focus on the original SiRF algorithm [33] without any repeated hold-out.

3.1 Simulation Setup

First, inspired by the correlation patterns of endocrine-disrupting chemicals from Midya et al. [38], we simulated correlated exposures with correlations varying from moderate to high correlation: 0.3 to 0.6; second, we generated exposure matrices with sample sizes 250 and 1000 and the number of exposures being 10, 25, 50, and, 100, respectively. Additionally, to mimic high-dimensional scenarios, exposure matrices with sample size 1000 and the number of exposures 250 and 500 were also generated; third, we assumed the simulated exposure one and exposure two had a positive linear association with the simulated outcome. On top of that, we created synergistic interactions of multiple orders similar to Basu et al. [33]. Assume V_1 ,V_2, ..., V_p, with p being the number of exposures; then we define the interactions as (1) order two: $I[V_1 > t_1 \& V_3 > t_3]$, (2) order three: $I[V_1 > t_1 \& V_3 > t_3 \& V_5 > t_5]$, and, (3) order four: $I[V_1 > t_1 \& V_3 > t_3 \& V_4 > t_4 \& V_5 > t_5]$, where $I(.)$ denotes an indicator function. The cutoffs were chosen to ensure a reasonable class balance ($\sim$ 30% with the synergistic interactions). All other exposures in respective simulations were kept inactive, i.e., no association. Thereafter, we created four different outcome scenarios under the assumption of additivity,

1. *Only main effect but no interaction*:
   Outcome $:=$ Main effects from V_1 and V_2 + Covariates + Gaussian error
2. *Main effect and two-order synergistic interaction*:
   Outcome $:=$ Main effects from V_1 and V_2 + $I[V_1 > t_1 \,\&\, V_3 > t_3]$ + Covariates + Gaussian error
3. *Main effect and three-order synergistic interaction*:
   Outcome $:=$ Main effects from V_1 and V_2 + $I[V_1 > t_1 \,\&\, V_3 > t_3 \,\&\, V_5 > t_5]$ + Covariates + Gaussian error
4. *Main effect and four-order synergistic interaction*:
   Outcome $:=$ Main effects from V_1 and V_2 + $I[V_1 > t_1 \,\&\, V_3 > t_3 \,\&\, V_4 > t_4 \,\&\, V_5 > t_5]$ + Covariates + Gaussian error.

Further simulations were conducted to ensure the Moxie algorithm could detect multiple interactions simultaneously (i.e., an outcome with two second-order interactions). Any dead or pseudo interaction was designated as a false positive. Any true interaction or smaller order interaction induced by it was adjudged as true positive since even smaller ordered features can be informative. Five different metrics were used to gauge the performances. Specificity was calculated for "Only main effect but no interaction." In contrast, sensitivity and false discovery rate (FDR) were calculated for the rest of the three scenarios. Lastly, we presented the recovery rate by interaction order; for example, if the outcome possessed three-ordered interactions, we asked whether the Moxie algorithm detects the three-ordered interaction, or can it only detect smaller ordered induced interactions or both.

3.2 Model Performance in Simulations

The performance of the Moxie algorithm was better than the SiRF in terms of specificity and the false discovery rate. Their performances were almost equivalent for sensitivity, but SiRF performed better for smaller sample sizes and higher order of interactions. Their performance for recovery rates was similar for smaller sample sizes; however, SiRF performed better for larger samples.

In most scenarios, the Moxie algorithm was less likely to detect false positives and more likely to choose true negatives. The specificity of the Moxie algorithm remained more than 95% in the entire simulated scenarios (Fig. 2), i.e., if there was no synergistic interaction, the Moxie algorithm was more likely to choose true negatives. Similarly, the FDR for the Moxie algorithm remained less than 5% irrespective of the sample size, number of exposures, or order of interaction (Fig. 3). SiRF was more likely to choose dead or pseudo interactions, and therefore, as the sample size increased, the specificity got worse. However, its FDR started to decrease when the sample size and the number of exposures substantially increased.

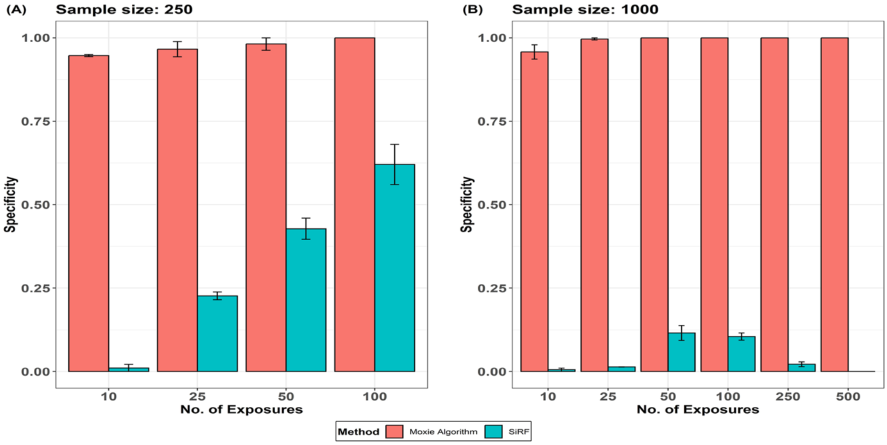

Fig. 2 Specificity (mean ± se) calculated for Moxie algorithm and SiRF in sample sizes **A** 250 and **B** 1000 while the number of exposures gradually increase

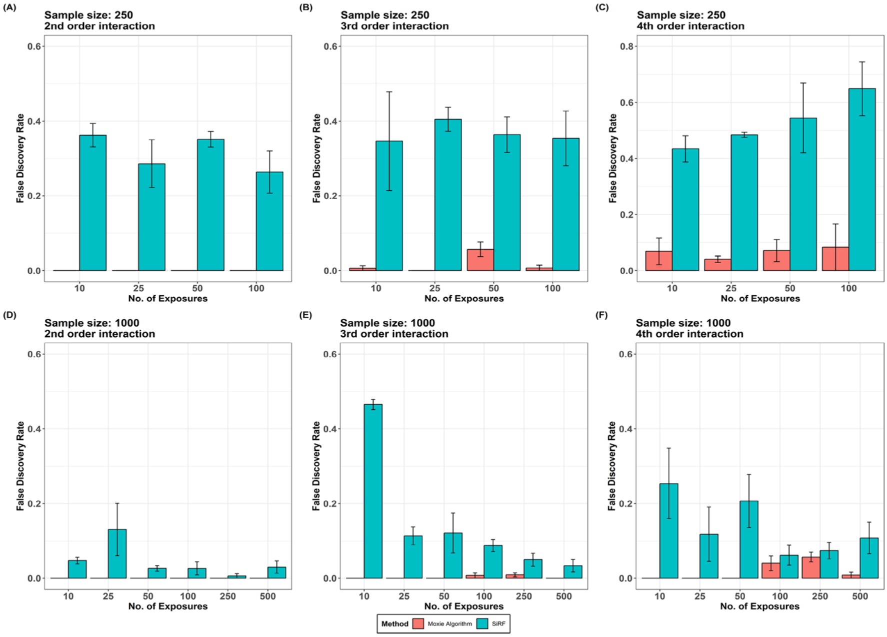

Fig. 3 False discovery rates (mean ± se) calculated for Moxie algorithm and SiRF in sample sizes **A** 250 and **B** 1000 while the number of exposures gradually increase

The sensitivity of the Moxie algorithm and SiRF was comparable (Fig. 4); However, the performance of SiRF was relatively better under a smaller sample size of 250 and a higher number of exposures, 50 and 100. When the sample size increased

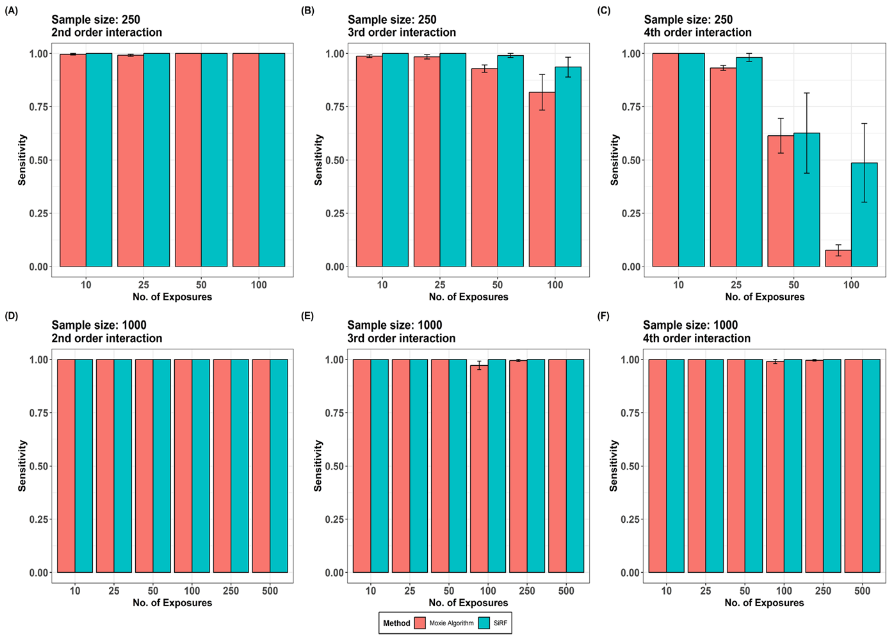

Fig. 4 Sensitivity (mean ± se) calculated for Moxie algorithm and SiRF in sample sizes **A** 250 and **B** 1000 while the number of exposures gradually increase

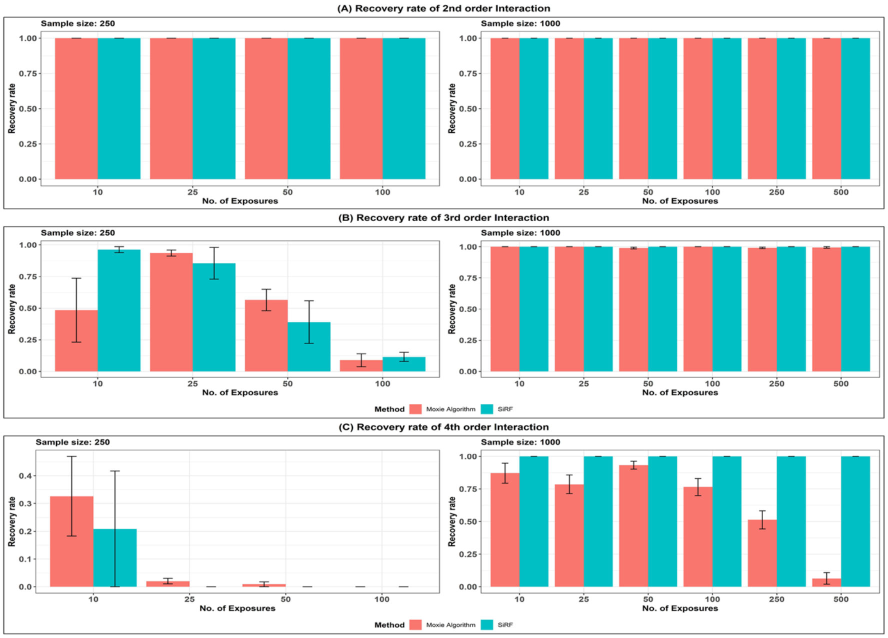

Fig. 5 Recovery rates (mean ± se) calculated for Moxie algorithm and SiRF for **A** 2nd-, **B** 3rd-, and **C** 4th-order interactions in sample sizes 250 and 1000 while the number of exposures gradually increase

to 1000, the sensitivities of both algorithms substantially increased, irrespective of the underlying order of interaction.

Finally, the recovery rate by interaction order is presented in Fig. 5. The Moxie algorithm and SiRF could efficiently recover any 2nd-order interaction. However, for 3rd- and 4th-order interactions, the recovery rates of both algorithms substantially decreased for smaller sample sizes (in these scenarios, both algorithms mostly recover lower-ordered induced interactions—keeping their sensitivity higher). For larger sample sizes and irrespective of the number of exposures, the recovery rates of SiRF remained at 100% for 3rd- and 4th-order interactions. However, for the Moxie algorithm, the recovery rate of 4th-order interaction decreased when the number of exposures substantially increased (> 250 for sample size $n = 1000$).

4 Application in US-NHANES 2017–18

4.1 Study Population

We used nationally representative and cross-sectional survey data from the 2017–2018 U.S. National Health and Nutrition Examination Survey (NHANES). The NHANES is conducted by the National Center for Health Statistics (NCHS) and the Centers for Disease Control and Prevention, and the analysis was conducted following their recommendations. Detailed descriptions of the NHANES survey design, data collection methodology, and analytical techniques can be found at CDC [39]. In this study, 447 adults, with ages ranging from 18 to 80 years, and participants with complete data on the outcome—low-density lipoprotein cholesterol (LDL-C), and the exposures—Polyfluoroalkyl substances (PFAS) and metal serum concentrations, were used for analysis.

4.2 Low-Density Lipoprotein Cholesterol as Outcome

LDL-C, known as 'bad cholesterol' and one of the major sources of cholesterol build-up and blockages in the arteries [40], is the outcome variable. LDL-C is a continuous outcome and was shown to be associated with serum PFAS and metal concentrations in previous literature [41–43], and [44], but lacked a search for potential interactions. All participants using prescription cholesterol-lowering statin drugs in this study are excluded from the analysis.

4.3 PFAS and Metals as Exposures

Serum PFAS included in the study were Perfluorodecanoic acid (PFDeA), Perfluorohexane sulfonic acid (PFHxS), 2-(N-methylperfluorooctanesulfonamido) acetic acid (Me-PFOSA-AcOH), Perfluorononanoic acid (PFNA), Perfluoroundecanoic acid (PFUA); Perfluorododecanoic acid (PFDoA), n-perfluorooctanoic acid (n-PFOA), Branch perfluorooctanoic acid isomers (Sb-PFOA), n-perfluorooctane sulfonic acid (n-PFOS), Perfluoromethylheptane sulfonic acid isomers (Sm-PFOS). All PFAS had at least 30% observations above the lower limit of detection. Metals included were Lead (Pb), Cadmium (Cd), Total Mercury (THg), Selenium (Se), and Manganese (Mn) and were measured in whole blood. All metals had at least 80% observations above LLOD. Values below LLODs were imputed by dividing the LLODs by 2 (Dong et al. 2019). Further details on the laboratory techniques used can be found [41]. See laboratory procedural manuals of 2017–2018 U.S. NHANES for further details.

4.4 Results

The demographic characteristics of this sample under study are presented in Table 1. Participants with high LDL-C ($\geq$ 130 mg/dL) [45] and [46] were more likely to be older, non-Hispanic black, have had alcohol in the past, were less likely to be physically active, and, had a higher concentration of all chemical exposures (except for Manganese). In the initial WQS model (with continuous outcome LDL-C and without interaction term), the mixture index was significantly associated with higher LDL-C levels (beta[95% CI]: 6.57[3.49, 9.64]). The chemicals Sm-PFOS, PFUA, Pb, Me-PFOSA-AcOH, and n-PFOA had higher contributions (weight $>$ 1/13) to the mixture index. The covariates in the model were selected based on prior literature [41] and [44].

Next, instead of directly extracting the residuals from this WQS model, we constructed a hypothetical experiment that mimics controlled randomized experiments to interpret better the chemical interaction term [47, 48] and [49]. First, the mixture index was dichotomized into a high vs. a low group based on its median value. Second, a matched-sampling strategy was used to obtain balance in covariate distributions between high-vs-low groups. A simple full matching with a caliper based on the estimated propensity score was used to construct similar groups with high and low mixture index having balanced covariate distribution [50] and [51]. Love plots of the differences in standardized means in covariates were used to examine whether covariate balancing was successful (setting the threshold for the standardized mean difference to 0.1) [52] and [53]. Given the covariates, we assumed that this approach of covariate balancing creates "exchangeable" groups so that the exposures were hypothetically and randomly assigned to each exposure-mixture group and ensured that the exposure assignment was not confounded by covariates. While conducting the covariate balancing, we adjusted the model for appropriate sampling weights from the 2017 to 2018 U.S. NHANES cycle. Note that using this causal-inference framework is not necessary but strengthens the interpretations in later stages while constructing counterfactual arguments. Finally, we extracted the residuals from a mixture model based on the matched sample and fitted the Moxie algorithm.

We found a two-order synergistic interaction between Cadmium (Cd) and Lead (Pb), denoted by Cd+/Pb+. Mechanistically, this interaction was expressed as a binary indicator where the interaction occurred when the whole blood concentrations of cadmium and lead were more than 0.605 ug/L and 1.485 ug/dL, respectively. Regarding directionality of association and statistical relevance, Cd+/Pb+

Table 1 Study characteristics of the population under investigation – data from National Health and Nutrition Examination Survey 2017-2018

	Total	LDL-C < 130 mg/dL	LDL-C ≥ 130 mg/dL	% observations ≥ LLOD	p-value
Sample size (n)	447	316	131		
Baseline Covariates					
Age (years)	44.23 (17.58)	41.34 (17.30)	51.21 (16.03)		<0.001
Ethnicity					0.89
Non-Hispanic Black	204 (45.64 %)	142 (44.94%)	62 (47.33%)		
Non-Hispanic White	136 (30.43%)	98 (31.01%)	38 (29.01%)		
Other Hispanic	107 (23.94%)	76 (24.05%)	31 (23.66%)		
Body mass index (kg/m^2)	29.64 (8.45)	29.26 (8.20)	30.55 (8.98)		
Gender					0.25
Male	223 (49.89 %)	152 (48.10%)	71 (54.20%)		
Female	224 (50.11 %)	164 (51.90%)	60 (45.80%)		
Smoking Status					0.01
Never	272 (60.85 %)	205 (64.87%)	67 (51.15%)		
Smoked at least 100 cigarettes in life					
but don’t smoke now	89 (19.91 %)	53 (16.77%)	36 (27.48 %)		
Smoked at least 100 cigarettes in life					
and still smoke now	86 (19.24 %)	58 (18.35%)	28 (21.37%)		
Ratio of family income to poverty	2.33 (1.55)	2.30 (1.53)	2.39 (1.60)		0.71
Alcohol intake: Ever had a drink of any kind of alcohol					0.86
Yes	404 (90.38 %)	286 (90.51%)	118 (90.08%)		
No	43 (9.62 %)	30 (9.49%)	13 (9.92%)		
Physical Activity: Vigorous work or recreational activity					<0.01
Yes	219 (48.99 %)	170 (53.80%)	49 (37.40%)		
No	228 (51.01 %)	146 (46.20%)	82 (62.60%)		

Table 1 (continued)

	Total	LDL-C < 130 mg/dL	LDL-C ≥ 130 mg/dL	% observations ≥ LLOD	p-value
PFAS exposures (Unadjusted geometric means with 95% confidence intervals)					
PFDeA (ng/mL)	0.19 (0.18, 0.21)	0.18 (0.17, 0.19)	0.22 (0.19, 0.25)	63.09 %	0.02
PFHxS (ng/mL)	1.02 (0.94, 1.10)	0.95 (0.86, 1.05)	1.21 (1.06, 1.39)	98.66%	<0.01
Me-PFOSA-AcOH (ng/mL)	0.12 (0.11, 0.13)	0.12 (0.11, 0.13)	0.13 (0.12, 0.15)	31.10%	0.08
PFNA (ng/mL)	0.37 (0.34, 0.40)	0.34 (0.31, 0.37)	0.44 (0.38, 0.51)	85.01%	<0.01
PFUA (ng/mL)	0.13 (0.13, 0.14)	0.12 (0.12, 0.13)	0.16 (0.14, 0.18)	38.03%	< 0.01
n-PFOA (ng/mL)	1.17 (1.10, 1.25)	1.11 (1.03, 1.19)	1.34 (1.21, 1.49)	99.55%	< 0.01
n-PFOS (ng/mL)	2.97 (2.73, 3.23)	2.74 (2.49, 3.02)	3.60 (3.04, 4.27)	99.78%	< 0.01
Sm-PFOS (ng/mL)	1.12 (1.03, 1.22)	1.01 (0.91, 1.11)	1.46 (1.26, 1.71)	98.66 %	< 0.01
Lead, Cadmium, Total Mercury, Selenium & Manganese exposures					
(*Unadjusted geometric means with 95% confidence intervals*)					
Cd (μg/L)	0.31 (0.29, 0.34)	0.30 (0.27, 0.33)	0.33 (0.28, 0.39)	90.38%	0.29
Pb (μg/dL)	0.84 (0.78, 0.90)	0.76 (0.70, 0.82)	1.05 (0.92, 1.19)	100%	< 0.01
Mn (μg/L)	9.60 (9.30, 9.92)	9.69 (9.32, 10.08)	9.39 (8.88, 9.93)	100%	0.36
THg (μg/L)	0.73 (0.67, 0.80)	0.69 (0.62, 0.76)	0.85 (0.71, 1.02)	83.22%	0.04
Se (μg/L)	188.00 (185.71, 190.32)	186.37 (183.68, 189.09)	191.99 (187.67, 196.41)	100%	0.03

Data presented as mean(SD) or *n*(%);*LLOD* lower limit of detection (in ng/mL), *LDL-C* low-density lipoprotein-cholesterol (mg/dL), *HDL-C* high-density lipoprotein-cholesterol (mg/dL), *PFDeA* Perfluorodecanoic acid, *PFHxS* Perfluorohexane sulfonic acid, *Me-PFOSA-AcOH* 2-(N-methylperfluorooctanesulfonamido)acetic acid, *PFNA* Perfluorononanoic acid, *PFUA* Perfluoroundecanoic acid, *PFDoA* Perfluorododecanoic acid, *n-PFOA* n-perfluorooctanoic acid, *Sb-PFOA* Branch perfluorooctanoic acid isomers, *n-PFOS* n-perfluorooctane sulfonic acid, *Sm-PFOS* Perfluoromethylheptane sulfonic acid isomers, *Pb* Lead, *Cd* Cadmium, *THg* Total Mercury, *Se* Selenium, *Mn* Manganese had a significantly positive association with increased LDL-C in a model with the main mixture index. Moreover, this interaction term improved the fit of the WQS regression (Likelihood ratio test statistic: 7.62, p value < 0.006). Cd+/Pb+ might also have potential biochemical significance. Metallothioneins (MTs) are cysteine-rich metal-binding proteins that bind to the biologically essential metals, perform homeostatic regulations of these metals, and absorb the heavy metals [54]. The MT2A core promoter region A/G (SNP) was shown to be associated with higher levels of Cd and Pb. In particular, individuals with the GG genotype were particularly more sensitive to heavy metal toxicity [54] and [55]. In a study with 221 car battery workers, blood lead levels were associated with genetic variation due to MT2A SNP [56]. Moreover, MTs possessed a strong binding affinity for Cadmium and were associated with oxidative effects and genotoxicity of cadmium [56]. On the other hand, MTs affect lipid metabolism by preventing lipid peroxidation and, therefore, might affect lipid profiles [57] and [58]. Further, [59] showed that the association between the blood lead level and serum lipid concentrations might be modified by the genetic combination of MT2A polymorphisms.

5 Concluding Remarks

In this paper, we presented the design and the utility of the Moxie algorithm—an amalgamation of Weighted Quantile Sum regression and interpretable machine-learning tools to extract and discover biologically plausible synergistic interactions among environmental chemicals. Most statistical tools devise interactions in terms of the projection of exposures and are usually difficult to interpret. Additionally, as the number of exposures increases, the computational demand to search for all possible combinations increases almost exponentially. However, with the help of interpretable shallow tree models, the Moxie algorithm bypasses most of these issues by focusing on the most predictive combinations, drastically decreasing the computational cost. Interactions obtained from this algorithm have the potential to be tested through in-vitro experiments for any biological plausibility. Moreover, these interactions can be thought to mimic toxicological interactions because of their construction. Through extensive simulations and real data examples, we demonstrated the use and applicability of this algorithm. Moxie algorithm can be deemed to be a “white-box” model with “trustworthy interpretations” [60], designed to discover toxicologically mimicking interactions, bridging a gap between difficult-to-comprehend machine-learning models with high-predictive power and classical mixture regression models with inferential prowess.

This algorithm also has limitations: (1) The Moxie algorithm was built to strictly limit false-positive interactions. While it performed well in specificity, sensitivity, and false discovery rates, its recovery of higher-order interaction broke down for smaller sample sizes while recovering larger orders of interactions. Therefore, a future direction would be to incorporate the agile and flexible nature of the SiRF algorithm while keeping the Moxie algorithm’s stringency. (2) This algorithm is currently limited to searching for synergistic interactions in adverse directions, but future developments can expand to synergistic and antagonistic interactions in both directions. (3) Lastly, this algorithm could disregard small but significant effect estimates due to its high quantile cutoffs. Nevertheless, provisions can be made to implement alternative measures of effect sizes, such as t-statistics, Cohen's d, and Likelihood Ratio Test statistics, to assimilate interactions with small but significant effect sizes.

This work demonstrates novelty in methodological development but also possesses much direct relevance to real-world problems. In this paper, we did not include any comparison with the repeated hold-out SiRF algorithm. Although it has been proposed in previous papers [37], selecting the top highly stable interactions is still subjective and context-specific, i.e., it is unclear how many of those top stable interactions can be considered "significant" and should be chosen for further downstream analysis. Future work can explore and validate such a mechanistic algorithm for selecting stable interactions. Moreover, depending on the context and the hypothesis, other exposure mixture tools, like the Bayesian Kernel Machine Regression and Quantile g-computation, can also be used in conjunction with the machine learning part of the Moxie algorithm. When the directionality of the association is hypothesized apriori, and the interest lies in evaluating uni-directional association, a WQS-based Moxie algorithm can be utilized (just as demonstrated in this paper). On the contrary, when the directionality of the association cannot be hypothesized apriori, or the interest lies in a measure of the overall association, a BKMR or Quantile g-computation-based Moxie algorithm can be utilized. In conclusion, we introduced a novel framework incorporating exposure mixture models and extreme gradient boosting techniques to discover toxicologically mimicking interactions. The amalgamation of such techniques provides ample opportunities for statistical method development and answering key problems about environmental health.

Data Availability The data are available online and has been pulled from NHANES 17-18.

Code Availability The codes for the MOXIE algorithm can be found in GitHub (https://github.com/vishalmidya/MOXIE).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Hamm AK, Hans Carter W Jr, Gennings C (2005) Analysis of an interaction threshold in a mixture of drugs and/or chemicals. Stat Med 24(16):2493–2507
2. Gibson EA (2021) Statistical and machine learning methods for pattern identification in environmental mixtures. Columbia University, New York
3. Gennings C (2000) On testing for drug/chemical interactions: definitions and inference. J Biopharm Stat 10(4):457–467
4. Gennings C, Carter W Jr, Carchman R, Teuschler L, Simmons J, Carney E (2005) A unifying concept for assessing toxicological interactions: changes in slope. Toxicol Sci 88(2):287–297
5. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P (2015) Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. J Agric Biol Environ Stat 20:100–120
6. Colicino E, Pedretti NF, Busgang SA, Gennings C (2020) Per-and poly-fluoroalkyl substances and bone mineral density: results from the bayesian weighted quantile sum regression. Environ Epidemiol 4(3):e092
7. Keil AP, Buckley JP, O'Brien KM, Ferguson KK, Zhao S, White AJ (2020) A quantile-based g-computation approach to addressing the effects of exposure mixtures. Environ Health Perspect 128(4):047004
8. Lee M, Rahbar MH, Samms-Vaughan M, Bressler J, Bach MA, Hessabi M, Grove ML, Shakespeare-Pellington S, Coore Desai C, Reece J-A et al (2019) A generalized weighted quantile sum approach for analyzing correlated data in the presence of interactions. Biom J 61(4):934–954
9. Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA (2015) Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics 16(3):493–508
10. Liu JZ, Deng W, Lee J, Lin P-ID, Valeri L, Christiani DC, Bellinger DC, Wright RO, Mazumdar MM, Coull BA (2022) A cross-validated ensemble approach to robust hypothesis testing of continuous non-linear interactions: application to nutrition-environment studies. J Am Stat Assoc 117(538):561–573
11. McGee G, Wilson A, Webster TF, Coull BA (2023) Bayesian multiple index models for environmental mixtures. Biometrics 79(1):462–474. https://doi.org/10.1111/biom.13569
12. Bellavia A (2021) Statistical methods for environmental mixtures. https://bookdown.org/andreabellavia/mixtures/preface.html. Accessed 10 Jan 2023
13. Bien J, Taylor J, Tibshirani R (2013) A lasso for hierarchical interactions. Ann Stat 41(3):1111
14. Gennings C, Schwartz P, Carter Jr WH, Simmons JE (1997) Detection of departures from additivity in mixtures of many chemicals with a threshold model. J Agric Biol Environ Stat, 2:198–211
15. Kelly C, Rice J (1990) Monotone smoothing with application to dose-response curves and the assessment of synergism. Biometrics 46:1071–1085
16. Machado SG, Robinson GA (1994) A direct, general approach based on isobolograms for assessing the joint action of drugs in pre-clinical experiments. Stat Med 13(22):2289–2309
17. Yeatts SD, Gennings C, Wagner ED, Simmons JE, Plewa MJ (2010) Detecting departure from additivity along a fixed-ratio mixture ray with a piecewise model for dose and interaction thresholds. J Agric Biol Environ Stat 15:510–522
18. Bhat AS, Ahangar AA (2007) Methods for detecting chemical-chemical interaction in toxicology. Toxicol Mech Methods 17(8):441–450
19. Shmueli G (2010) To explain or to predict? Stat Sci 25(3):289–310
20. Gass K, Klein M, Chang HH, Flanders WD, Strickland MJ (2014) Classification and regression trees for epidemiologic research: an air pollution example. Environ Health 13(1):1–10
21. Lampa E, Lind L, Lind P, Bornefalk-Hermansson A (2014) The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. Environ Health 13:57
22. Li Y-C, Hsu H-HL, Chun Y, Chiu P-H, Arditi Z, Claudio L, Pandey G, Bunyavanich S, et al. (2021) Machine learning–driven identification of early-life air toxic combinations associated with childhood asthma outcomes. J Clin Investig 131(22):e152088
23. Stingone JA, Pandey OP, Claudio L, Pandey G (2017) Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among us children. Environ Pollut 230:730–740
24. Chen T, Guestrin C (2016) XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM

25. Curtin P, Kellogg J, Cech N, Gennings C (2021) A random subset implementation of weighted quantile sum (wqsrs) regression for analysis of high-dimensional mixtures. Commun Stat Simul Comput 50(4):1119–1134
26. Tanner EM, Bornehag C-G, Gennings C (2019) Repeated holdout validation for weighted quantile sum regression. MethodsX 6:2855–2860
27. Joubert BR, Kioumourtzoglou M-A, Chamberlain T, Chen HY, Gennings C, Turyk ME, Miranda ML, Webster TF, Ensor KB, Dunson DB et al (2022) Powering research through innovative methods for mixtures in epidemiology (prime) program: novel and expanded statistical methods. Int J Environ Res Public Health 19(3):1378
28. Biau G, Scornet E (2016) A random forest guided tour. TEST 25:197–227
29. Gelfand S, Ravishankar C, Delp E (1991) An iterative growing and pruning algorithm for classification tree design. IEEE Trans Pattern Anal Mach Intell 13(2):163–174
30. Lin J (2008) Scalable language processing algorithms for the masses: a case study in computing word co-occurrence matrices with MapReduce. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 419–428, Honolulu, Hawaii. Association for Computational Linguistics
31. Li Y, Xu L, Tian F, Jiang L, Zhong X, Chen E (2015) Word embedding revisited: a new representation learning and explicit matrix factorization perspective. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, page 3650-3656. AAAI Press
32. Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. Ann Appl Stat 2(3):916–954
33. Basu S, Kumbier K, Brown JB, Yu B (2018) Iterative random forests to discover predictive and stable high-order interactions. Proc Natl Acad Sci 115(8):1943–1948
34. Kumbier K, Basu S, Brown JB, Celniker S, Yu B (2018) Refining interaction search through signed iterative random forests. arXiv preprint arXiv:1810.07287
35. Shah RD, Meinshausen N (2014) Random intersection trees. J Mach Learn Res 15(1):629–654
36. Midya V, Alcala CS, Rechtman E, Gregory JK, Kannan K, Hertz-Picciotto I, Teitelbaum SL, Gennings C, Rosa MJ, Valvi D (2023a) Machine learning assisted discovery of interactions between pesticides, phthalates, phenols, and trace elements in child neurodevelopment. Environ Sci Technol 57(46):18139–18150. https://doi.org/10.1021/acs.est.3c00848
37. Midya V, Lane JM, Gennings C, Torres-Olascoaga LA, Gregory JK, Wright RO, Arora M, Téllez-Rojo MM, Eggers S (2023b) Prenatal lead exposure is associated with reduced abundance of beneficial gut microbial cliques in late childhood: an investigation using microbial co-occurrence analysis (MiCA). Environ Sci Technol 57(44):16800–16810. https://doi.org/10.1021/acs.est.3c04346
38. Midya V, Colicino E, Conti DV, Berhane K, Garcia E, Stratakis N, Andrusaityte S, Basagaña X, Casas M, Fossati S, Gražulevičienė R, Haug LS, Heude B, Maitre L, McEachan R, Papadopoulou E, Roumeliotaki T, Philippat C, Thomsen C, Urquiza J, Vafeiadi M, Varo N, Vos MB, Wright J, McConnell R, Vrijheid M, Chatzi L, Valvi D (2022) Association of prenatal exposure to endocrine-disrupting chemicals with liver injury in children. JAMA Netw Open 5(7):e2220176–e2220176
39. CDC U (2013) Fourth national report on human exposure to environmental chemicals, updated tables. CDC, U
40. Dong Z, Wang H, Yu YY, Li YB, Naidu R, Liu Y (2019) Using 2003–2014 us nhanes data to determine the associations between per-and polyfluoroalkyl substances and cholesterol: trend and implications. Ecotoxicol Environ Saf 173:461–468
41. Buhari O, Dayyab F, Igbinoba O, Atanda A, Medhane F, Faillace R (2020) The association between heavy metal and serum cholesterol levels in the us population: National health and nutrition examination survey 2009–2012. Hum Exp Toxicol 39(3):355–364
42. Jain RB, Ducatman A (2018) Associations between lipid/lipoprotein levels and perfluoroalkyl substances among us children aged 6–11 years. Environ Pollut 243:1–8
43. Liu HS, Wen LL, Chu PL, Lin CY (2018) Association among total serum isomers of perfluorinated chemicals, glucose homeostasis, lipid profiles, serum protein and metabolic syndrome in adults: NHANES, 2013–2014. Environ Pollut 232:73–79
44. Midya V, Liao J, Gennings C, Colicino E, Teitelbaum SL, Wright RO, Valvi D (2022) Quantifying the effect size of exposure-outcome association using δ-score: application to environmental chemical mixture studies. Symmetry 14(10):1962
45. Fernández-Friera L, Fuster V, López-Melgar B, Oliva B, García-Ruiz JM, Mendiguren J, Bueno H, Pocock S, Ibáñez B, Fernández-Ortiz A et al (2017) Normal ldl-cholesterol levels are associated with subclinical atherosclerosis in the absence of risk factors. J Am Coll Cardiol 70(24):2979–2991

46. Jellinger PS, Handelsman Y, Rosenblit PD, Bloomgarden ZT, Fonseca VA, Garber AJ, Grunberger G, Guerin CK, Bell DS, Mechanick JI et al (2017) American association of clinical endocrinologists and American college of endocrinology guidelines for management of dyslipidemia and prevention of cardiovascular disease. Endocr Pract 23:1–87
47. Bind M-AC, Rubin DB (2019) Bridging observational studies and randomized experiments by embedding the former in the latter. Stat Methods Med Res 28(7):1958–1978
48. Rubin DB (2008) For objective causal inference, design trumps analysis. Ann Appl Stat 2(3):808–840. https://doi.org/10.1214/08-AOAS187
49. Sommer AJ, Peters A, Rommel M, Cyrys J, Grallert H, Haller D, Müller CL, Bind M-AC (2022) A randomization-based causal inference framework for uncovering environmental exposure effects on human gut microbiota. PLoS Comput Biol 18(5):e1010044
50. Hansen BB (2004) Full matching in an observational study of coaching for the sat. J Am Stat Assoc 99(467):609–618
51. Ho D, Imai K, King G, Stuart EA (2011) MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 42(8):1–28. https://doi.org/10.18637/jss.v042.i08
52. Greifer N (2020) Covariate balance tables and plots: a guide to the cobalt package. Accessed 10 Mar 2020
53. Zhang Z, Kim HJ, Lonjon G, Zhu Y et al (2019) Balance diagnostics after propensity score matching. Ann Transl Med 7(1):16
54. Kayaaltı Z, Aliyev V, Söylemezoğlu T (2011) The potential effect of metallothionein 2A–5 A/G single nucleotide polymorphism on blood cadmium, lead, zinc and copper levels. Toxicol Appl Pharmacol 256(1):1–7
55. Verma N, Bal S, Gupta R, Aggarwal N, Yadav A (2020) Antioxidative effects of piperine against cadmium-induced oxidative stress in cultured human peripheral blood lymphocytes. J Diet Suppl 17(1):41–52
56. Fernandes KCM, Martins AC Jr, Oliveira AÁSd, Antunes LMG, Cólus IMdS, Barbosa F Jr, Barcelos GRM (2016) Polymorphism of metallothionein 2a modifies lead body burden in workers chronically exposed to the metal. Public Health Genomics 19(1):47–52
57. Yang X, Sun J, Ke H, Chen Y, Xu M, Luo G (2014) Metallothionein 2a genetic polymorphism and its correlation to coronary heart disease. Eur Rev Med Pharmacol Sci 18:3747–3753
58. Ling X-B, Wei H-W, Wang J, Kong Y-Q, Wu Y-Y, Guo J-L, Li T-F, Li J-K (2016) Mammalian metallothionein-2a and oxidative stress. Int J Mol Sci 17(9):1483
59. Yang C-C, Chuang C-S, Lin C-I, Wang C-L, Huang Y-C, Chuang H-Y (2017) The association of the blood lead level and serum lipid concentrations may be modified by the genetic combination of the metallothionein 2a polymorphisms rs10636 gc and rs28366003 aa. J Clin Lipidol 11(1):234–241
60. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci 116(44):22071–22080

Authors and Affiliations

Vishal Midya[1] · **Chris Gennings**[1]

✉ Vishal Midya
vishal.midya@mssm.edu

[1] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York City, NY, USA