Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2013 *Paper* 102

A Bayesian regression tree approach to identify the effect of nanoparticles properties on toxicity profiles

Cecile Low-Kam[*] Haiyuan Zhang[†] Zhaoxia Ji[‡]
Tian Xia[**] Jeffrey I. Zinc[††]
Andre Nel[‡‡] Donatello Telesca[§]

[*]UCLA Biostatistics - CNSI, clowkam@ucla.edu
[†]CNSI
[‡]CNSI
[**]UCLA Nanomedicine - CNSI
[††]UCLA Chemistry and Biochemistry
[‡‡]UCLA Nanomedicine - CNSI
[§]UCLA, Biostatistics, dtelesca@ucla.edu

A Bayesian regression tree approach to identify the effect of nanoparticles properties on toxicity profiles

Cecile Low-Kam, Haiyuan Zhang, Zhaoxia Ji, Tian Xia, Jeffrey I. Zinc, Andre Nel, and Donatello Telesca

Abstract

We introduce a Bayesian multiple regression tree model to characterize relationships between physico-chemical properties of nanoparticles and their in-vitro toxicity over multiple doses and times of exposure. Unlike conventional models that rely on data summaries, our model solves the low sample size issue and avoids arbitrary loss of information by combining all measurements from a general exposure experiment across doses, times of exposure, and replicates. The proposed technique integrates Bayesian trees for modeling threshold effects and interactions, and penalized B-splines for dose and time-response surfaces smoothing. The resulting posterior distribution is sampled via a Markov Chain Monte Carlo algorithm. This method allows for inference on a number of quantities of potential interest to substantive nanotoxicology, such as the importance of physico-chemical properties and their marginal effect on toxicity. We illustrate the application of our method to the analysis of a library of 24 nano metal oxides.

A Bayesian regression tree approach to identify the effect of nanoparticles properties on toxicity profiles

CECILE LOW-KAM[1,2], HAIYUAN ZHANG[2], ZHAOXIA JI[2],
TIAN XIA[2,3], JEFFREY I. ZINK[2,4],
ANDRE E. NEL[2,4], DONATELLO TELESCA[1,2]

Author's Footnote

[1] UCLA Fielding School of Public Health, Department of Biostatistics
[2] UCLA California NanoSystems Institute
[3] UCLA Department of Medicine, Division of NanoMedicine
[4] UCLA Department of Chemistry and Biochemistry

March 2, 2013

Abstract

We introduce a Bayesian multiple regression tree model to characterize relationships between physico-chemical properties of nanoparticles and their in-vitro toxicity over multiple doses and times of exposure. Unlike conventional models that rely on data summaries, our model solves the low sample size issue and avoids arbitrary loss of information by combining all measurements from a general exposure experiment across doses, times of exposure, and replicates. The proposed technique integrates Bayesian trees for modeling threshold effects and interactions, and penalized B-splines for dose and time-response surfaces smoothing. The resulting posterior distribution is sampled via a Markov Chain Monte Carlo algorithm. This method allows for inference on a number of quantities of potential interest to substantive nanotoxicology, such as the importance of physico-chemical properties and their marginal effect on toxicity. We illustrate the application of our method to the analysis of a library of 24 nano metal oxides.

Keywords: Bayesian CART; Nanotoxicology; P-Splines; Regression trees.

1 Introduction

The increasing use of engineered nanomaterials (ENM) in hundreds of consumer products has recently risen concern about their potential effect on the environment and human health in particular. In nanotoxicology, in vitro dose-escalation assays describe how cell lines or simple organisms are affected by increased exposure to nanoparticles. These assays help determine hazardous materials and exposure levels. Standard dose-escalation studies are sometimes completed by more general exposure escalation protocols, where a biological outcome is measured against both increasing concentrations and durations of exposure.

Cost and timing issues often only allow for a small number of particles to be comprehensively screened in any study. Therefore, both one and two-dimensional escalation experiments are often characterized by small sample size, intended as number of particles. Furthermore, data often exhibits natural clusters related to varying levels of particles bio-activity. The two case studies presented in Section 6 provide an overview of the structure of typical datasets obtained with both experimental protocols.

Beyond dose response analysis, nanomaterial libraries are often designed to investigate how a range of physical and chemical properties (size, shape, composition, surface characteristics) may influence ENM's interactions with biological systems. The nano-informatics literature reports several Quantitative Structure-Activity Relationship (QSAR) models. This exercise is often conceived as a framework for predictive toxicology, under the assumption that nanoparticles with similar properties are likely to have similar effects. Most of existing QSAR models summarize or integrate experimental data across times, doses and replicates as a preprocessing step, before applying traditional data mining or statistical algorithms for regression. For example, Liu et al. (2011) use a modified Student's t-statistic to discretize outputs in two classes (toxic or non-toxic) and a logistic regression model. Zhang et al. (2012) use the area under the dose-response curve as a global summary of toxicity and they model dependence on predictors via a regression tree. Both approaches, while reasonably sensible, ignore the uncertainty associated with data summaries and can lead to unwarranted conclusions as well as unnecessary loss of information. On the other hand, the use of regression trees is inherently appealing as they are able to model non-linear effects and interactions with adaptive parsimony, without compromising interpretation. We aim to extend these regression

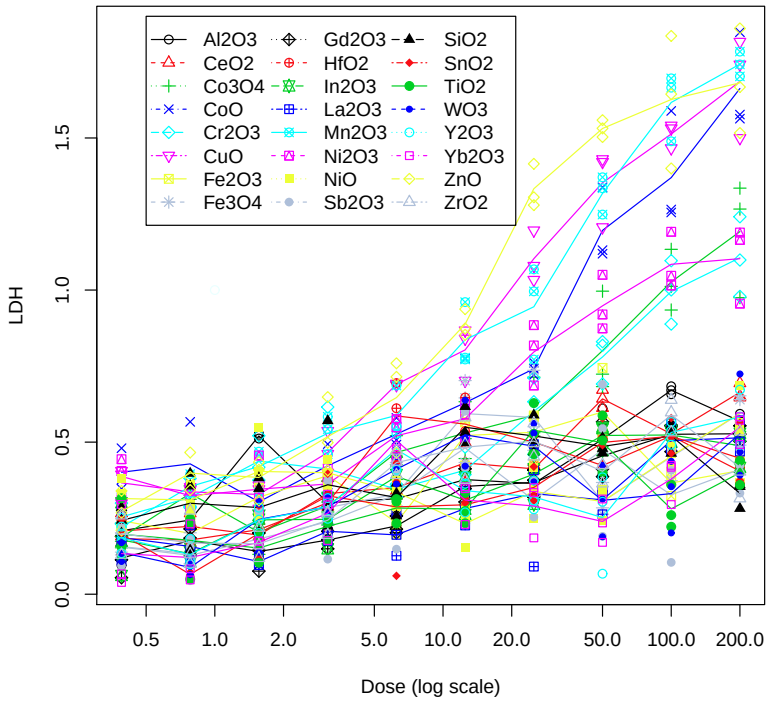

Figure 1: Dose-response curves for LDH assay.

models to account for structured multivariate outcomes, defined as toxicity profiles of nanoparticles, measured over a general exposure escalation domain.

Multivariate extensions of the regression tree methodology have been proposed by Segal (1992). In this paper, the original tree-building algorithm of Breiman et al. (1984) is modified to handle multivariate responses for commonly used covariance matrices, such as independence or autoregressive structures. Segal and Xiao (2011) extend this last approach to random forests, which combine multiple trees. De'ath (2002) proposes a similar method for an independent covariance structure. Yu and Lambert (1999) develop regression tree models for functional data, by representing each individual response as a linear combination of spline basis functions, and then proceeding with the estimated coefficients as new multivariate data to perform multivariate regression trees.

An alternative for longitudinal responses consists in combining a tree model and a linear model: Sela and Simonoff (2012) replace the fixed effects of the traditional linear mixed effects model by

a regression tree. The linear random effects are unchanged. Yu et al. (2010) fit a semi-parametric model, containing a linear part and a tree part, for multivariate outcomes in genetics. The linear part is used to model main effects of some genetic or environmental exposures. The non-parametric tree part approximates the joint effect of these exposures. Finally, Galimberti and Montanari (2002) develop regression tree models for longitudinal data with time-dependent covariates. In this setting, measures for a same individual can belong to different terminal nodes.

Other extensions of standard regression trees include Bayesian approaches, where tree parameters become random variables. Chipman et al. (1997) introduce a Bayesian regression tree model for univariate responses. The method is based on a prior distribution and a Metropolis-Hastings algorithm which generates candidate trees and identifies the most promising ones. This methodology has since been extended to so-called *treed* models, where a parametric model is fitted in each terminal node (Chipman et al. (2002)), for multivariate responses (Pang (2009)), to a sum-of-trees model (Chipman et al. (2010)), and to incorporate spatial random effects for merging datasets (Müller et al. (2007)), among others.

Building on previous contributions, we propose a new method to analyze the relationship between nanoparticles physio-chemical properties and their toxicity in exposure escalation experiments. We extend the Bayesian methodology of Chipman et al. (1997) to allow for dose and time response kinetics in terminal nodes. A global covariance structure accounts for correlation between measurements at different doses and times for a same particle. Our approach is able to model non-linear effects and potential interactions of physio-chemical properties without making parametric assumptions about toxicity profiles. It also addresses the issues associated with conventional QSAR models by combining evidence across measurements for all doses and times in a general experimental design. . The proposed model is particularly versatile as it provides scores of importance for physio-chemical properties, and visual assessment of the marginal effect of these properties on toxicity.

The rest of this paper is organized as follows: Section 2 describes the regression model for dose-response data and Section 3 the corresponding prior model. The resulting posterior distribution and the associated MCMC algorithm are presented in Section 4. The model is extended to the case of dose and time response surfaces in Section 5. The method is applied to a library of 24 metal oxides in Section 6. Finally, Section 7 concludes this paper and discuss its possible extensions.

2 Regression Tree Formulation

2.1 Sampling Model

We first consider the case of a typical dose escalation experiment, where a biological outcome is measured over a protocol of increased particle concentration. This case will be expanded in Section 5 to include more general exposure escalation designs.

Let $y_{ik}(d)$ denote a real valued response associated with exposure to nanoparticle i and replicate k at dose d, for $i \in \{1, ..., I\}$, $k \in \{1, ..., K\}$ and $d \in [0, D]$. We assume that y has been appropriately normalized and purified of experimental artifacts. Current experimental protocols only allow for the observation of the outcome y as it varies in association with a discrete prescription of dose-escalation. However, for notational convenience and without loss of generality, we maintain that y shall be observed for any dose level d ranging from no exposure ($d = 0$) to a maximal particle concentration level ($d = D$). Let also $\boldsymbol{x}_i' = (x_{i1}, ..., x_{ip})$ be a p-dimensional vector of continuous physio-chemical characteristics or predictors associated to particle i. We assume

$$y_{ik}(d) \quad = \quad f(\boldsymbol{x}_i, d) + \varepsilon_{ik}(d); \tag{1}$$

where f is a random mean function, depending on the dose level d and particle characteristics $\boldsymbol{x}_i$, and $\varepsilon_{ik} \sim N(0, \sigma_d^2)$.

More precisely, f is defined by a regression tree $\mathcal{T}$ on the predictor space, and a functional model for dose-response curves in the terminal nodes of $\mathcal{T}$. Full details about the proposed mean structure are described in the following section.

Given f, we assume that outcomes are independent across particles and, for any particle i, we assume

$$\mathrm{Cov}\big(\varepsilon_{ik}(d), \varepsilon_{ik'}(d')\big) \quad = \quad \sigma^2 \varphi_D^{|d-d'|}, \tag{2}$$

with $\varphi_D \in [0, 1]$. In this setting, two outcomes associated with the same particle at similar doses are assumed to be more correlated than measurements taken at distant doses, for any replicate. The major advantage of this assumption is related to a reduced representation of a high dimensional covariance matrix, which is now fully characterized in terms of a 1-dimensional variance parameter σ^2 and a 1-dimensional correlation φ_D.

2.2 Mean Structure

The binary tree $\mathcal{T}$ recursively splits the predictor space into two subspaces, according to criteria of the form $x_{\cdot j} \leq a$ vs $x_{\cdot j} > a$, for $a \in \mathbb{R}$ and $j \in \{1, ..., p\}$. Each split defines two new nodes of the tree, corresponding to two newly created subspaces of predictors. Let n be the set of terminal nodes of tree $\mathcal{T}$.

We model the dose-response curves in each terminal node as a linear combination of spline basis functions. Unlike parametric models such as log-logistic, spline functions do not assume a particular shape for the curve. This makes our model fully applicable to sub-lethal biological assays, which are not expected to follow a sigmoidal dose-response dynamic. Moreover, this functional structure is easily extended to two-dimensional response surfaces (Section 5). Let $\mathcal{B}_1(.), ..., \mathcal{B}_{m_D+\delta}(.)$ denote $m_D + \delta$ uniform B-spline basis functions of order δ on $[0, D]$, with m_D fixed knots. Following Eilers and Marx (1996), we avoid choosing the location of spline interior knots by deliberately overfitting curves with a number of knots coinciding with the dose-escalation design grid. Adaptive smoothness is determined by using a penalty on adjacent coefficients, via a smoothing prior that will be presented in Section 3.

If $\boldsymbol{x}_i$ is in the subset corresponding to the r^{th} terminal node of $\mathcal{T}$,

$$f(\boldsymbol{x}_i, d) \quad = \quad \sum_{\ell=1}^{m_D+\delta} \beta_{r\ell} \mathcal{B}_\ell(d). \tag{3}$$

We will denote with $\boldsymbol{\beta}_r = (\beta_{r1}, ..., \beta_{r\, m_D+\delta})'$, the vector of splines coefficients defining the expected dose-response trajectory in the r^{th} terminal node. Furthermore we let $\boldsymbol{\beta}$ define the random set of spline coefficients, including $\boldsymbol{\beta}_r$ from all terminal nodes ($r = 1, \ldots, n$).

The assumed covariance structure, in addition to tree splits and terminal nodes parameters, entirely define the distribution of the responses $y_{ik}(d)$. The Bayesian model is completed by prior distributions on $\mathcal{T}$, $\boldsymbol{\beta}$, σ^2 and φ_D.

3 Prior Model

We first introduce the general dependence structure of the prior, before describing each parameter's prior distribution. We follow Chipman et al. (2010), and assume that the tree is independent of

variance components σ^2 and φ_D:

$$\begin{aligned} p(\mathcal{T}, \boldsymbol{\beta}, \sigma^2, \varphi_D) &= p(\mathcal{T}, \boldsymbol{\beta})\, p(\sigma^2)\, p(\varphi_D) \quad (4) \\ &= p(\boldsymbol{\beta} \mid \mathcal{T}) p(\mathcal{T})\, p(\sigma^2)\, p(\varphi_D), \quad (5) \end{aligned}$$

Moreover, conditionally on $\mathcal{T}$, terminal node parameters are assumed independent:

$$p(\boldsymbol{\beta} \mid T) = \prod_{r=1}^{n} p(\boldsymbol{\beta}_r \mid \mathcal{T}). \quad (6)$$

Therefore, the prior is fully determined by a tree prior $p(\mathcal{T})$, terminal node parameters priors $p(\boldsymbol{\beta}_r \mid \mathcal{T})$ and variance parameters priors $p(\sigma^2)$ and $p(\varphi_D)$.

3.1 Tree Prior

The tree prior $p(\mathcal{T})$ is implicitly described by the stochastic tree-generating process of Chipman et al. (1997), where each new tree is generated according to:

i. the probability for a node at depth q to be non-terminal, given by $\alpha(1+q)^{-\nu}$, $(q = 1, 2, \ldots)$,

ii. the probability for a node to split at a predictor $x_{\cdot j}$, $(j = 1, \ldots, p)$, given by the discrete uniform distribution on the set of available predictors,

iii. given the predictor $x_{\cdot j}$, the probability for a node to split at a value a, given by the discrete uniform distribution on the set of available splitting values.

Probability i. is a decreasing function of q, making deeper nodes less likely to split and favoring "bushy" trees. Chipman et al. (1997) give guidelines to choose parameters α and ν by plotting the marginal prior distribution of the number of terminal nodes. In ii. and iii., predictors and splits are available if they lead to non-empty child nodes.

3.2 Terminal Node Splines Coefficients Prior

We follow Lang and Brezger (2004) and consider a conditionally conjugate P-spline prior:

$$\boldsymbol{\beta}_r \mid \mathcal{T}, \tau^2 \;\propto\; \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}_r' K_\beta \boldsymbol{\beta}_r\right), \quad (7)$$

where τ^2 is an additional smoothing variance parameter and

$$K_\beta = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \tag{8}$$

is a penalty matrix of size $(m_D + \delta) \times (m_D + \delta)$, corresponding to a first order random walk. Note that this prior is improper as matrix K_β is not of full rank. To obtain a proper prior and enable comparisons between trees of different sizes, in practice, we replace the first and last element of the diagonal with $1 + \eta$, where η is a small constant.

The model is completed by assigning a conjugate Inverse-Gamma hyperprior to the smoothing parameter $\tau^2 \mid \mathcal{T} \sim IG(a_\tau, b_\tau)$.

3.3 Variance Components Priors

We assume $\sigma^2 \sim IG(a_\sigma, b_\sigma)$. For φ_D, we choose the conjugate prior described in (Rowe (2003)) for autoregressive covariance matrices, with truncated support on $[0, 1]$. Let $0 = d_1 < ... < d_{n_D} = D$ be the dose-escalation design sequence:

$$p(\varphi_D) \propto \left(1 - \varphi_D^2\right)^{-\frac{n_D - 1}{2}} \exp\left(-\frac{\lambda_{01} - \varphi_D \lambda_{02} + \varphi_D^2 \lambda_{03}}{2\left(1 - \varphi_D^2\right)}\right) \mathbb{I}\{\varphi_D \in [0, 1]\}, \tag{9}$$

where $\mathbb{I}$ is the indicator function, $\Lambda = (\Lambda_{vv'})_{1 \leq v, v' \leq n_D}$ is a hyperparameter matrix, and $(\lambda_{01}, \lambda_{02}, \lambda_{03})$ are defined through its diagonal, subdiagonal and superdiagonal elements as:

$$\lambda_{01} = \sum_{v=1}^{n_D} \Lambda_{vv}, \tag{10}$$

$$\lambda_{02} = \sum_{v=1}^{n_D - 1} \left(\Lambda_{vv+1} + \Lambda_{v+1v}\right), \tag{11}$$

and

$$\lambda_{03} = \sum_{d=2}^{n_D - 1} \Lambda_{vv}. \tag{12}$$

In practice, we choose $\Lambda = Id_{n_D}$, the identity matrix of size $n_D \times n_D$, to put more weight on low values of φ_D and assume weak prior correlations between responses at different doses. This last distribution completes the prior model. We now turn to posterior inference on parameters, given the observations.

4 Posterior inference through MCMC simulation

We are interested in the posterior distribution

$$p\left(\mathcal{T}, \boldsymbol{\beta}, \sigma^2, \varphi_D, \tau^2 \mid \boldsymbol{y}\right). \tag{13}$$

The rest of this section describes a Markov chain Monte Carlo algorithm to sample from this distribution, as the number of potential trees prevents direct calculations. The algorithm is adapted from the Gibbs sampler of Chipman et al. (2010), with changes due to the specific structure of our model.

At each iteration, the algorithm performs a joint update of ($\mathcal{T}$, $\boldsymbol{\beta}$), conditionally on the rest of the parameters, followed by standard Gibbs component-wise updates of each variance parameter. The joint tree and terminal nodes splines coefficients update is decomposed into

$$\mathcal{T} \quad | \quad \boldsymbol{y}, \sigma^2, \varphi_D, \tau^2; \tag{14}$$

followed by

$$\boldsymbol{\beta}_r \quad | \quad \mathcal{T}, \boldsymbol{y}, \sigma^2, \varphi_D, \tau^2; \tag{15}$$

for $r \in \{1, ..., n\}$.

The draw of $\mathcal{T}$ in (14) is performed by the Metropolis-Hastings algorithm of Chipman et al. (1997), which simulates a Markov chain of trees that converges to the posterior distribution $p(\mathcal{T} \mid \boldsymbol{y}, \sigma^2, \varphi_D, \tau^2)$. The proposal density suggests a new tree based on four moves: grow a terminal node, prune a pair of terminal nodes, change the split rule of an internal node, and swap the splits of an internal node and one of its children's. To improve mixing, Wu et al. (2007) add a new move that allows for larger changes in trees but leaves the partition of observations in terminal nodes unchanged.

The target distribution can be decomposed as follows:

$$p(\mathcal{T} \mid \boldsymbol{y}, \sigma^2, \varphi_D, \tau^2) \quad \propto \quad p(\mathcal{T}) \int p(\boldsymbol{y} \mid \boldsymbol{\beta}, \mathcal{T}, \sigma^2, \varphi_D, \tau^2) p(\boldsymbol{\beta} \mid \mathcal{T}, \sigma^2, \varphi_D, \tau^2) d\boldsymbol{\beta}. \tag{16}$$

The full expression for the integral in the expression above is given in Equation (B.23) of Appendix **??** of Supplementary Material, and is obtained in a closed form by conjugacy of the prior on

$\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_n\}$. Therefore, the draw of $\mathcal{T}$ in (14) does not require a reversible-jump procedure for spaces of varying dimensions, even if nodes are added or deleted. The proposal density of the Metropolis-Hastings algorithm can be conveniently coupled with $p(\mathcal{T})$ to simplify calculations (Chipman et al. (1997)). Full conditional distributions for $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_n$ in (15) and variance parameters σ^2, φ_D and τ^2 are also available in supplementary Appendix **??**.

Given posterior samples, predictive statistics are easily obtained via Monte Carlo simulation of $p(\boldsymbol{y}_i^* \mid \boldsymbol{y})$, for $i = 1, \ldots, I$. More precisely, let $\boldsymbol{x}_i^* = \boldsymbol{x}_i$. At each iteration $\ell = 1, ..., N$, the MCMC algorithm performs a draw from $p\left(\mathcal{T}, \boldsymbol{\beta}, \sigma^2, \varphi_D, \tau^2 \mid \boldsymbol{y}\right)$, followed by a draw of $\boldsymbol{y}_i^{(\ell)*}$ from $p(\boldsymbol{y}_i^{(\ell)*} | \mathcal{T}, \boldsymbol{\beta}, \sigma^2, \varphi_D, \tau^2)$, a multivariate normal distribution. In our case studies (§6), for example, we compare posterior summaries from the predictive distribution $p(y_{ik}^*(d) \mid \boldsymbol{y})$ to observed dose-response data $y_{ik}(d)$, to assess model adequacy and calibration.

Posterior inference based on Monte Carlo samples, is also used to derive inferential summaries about non-trivial functionals of the parameter/model space. Let $\mathcal{T}^{(1)}, ..., \mathcal{T}^{(N)}$ be the regression trees generated by the MCMC algorithm. For all $j = 1, \ldots, p$ and $\ell = 1, \ldots, N$, let $z_j^{(\ell)}$ be the number of splits of tree $\mathcal{T}^{(\ell)}$ using variable $x_{.j}$. Then $\sum_{\ell=1}^{N} z_j^{(\ell)}$ gives an importance score for physio-chemical property $x_{.j}$ (Chipman et al. (2010)). These scores can be used as a preliminary step for variable selection, for a better structural interpretation of the impact of each physio-chemical property on toxicity.

Similarly, the marginal effect of a physio-chemical property $x_{.j}$ can be represented by the partial dependence function of Friedman (2001): let $x_{1j}, ..., x_{Sj}$ be a grid of new values for $x_{.j}$. Then the partial dependence function is defined at all $s = 1, ..., S$ as

$$f(x_{sj}, d, t) \quad = \quad \frac{1}{I} \sum_{i=1}^{I} f\left(\left(x_{i1}, ..., x_{ij-1}, x_{sj}, x_{ij+1}, ..., x_{ip}\right), d, t\right), \tag{17}$$

where $x_{ij'}$ is the i^{th} observation of $x_{.j'}$ in the data. For all doses, the average of this function over Monte Carlo draws, is then used as an estimate of the marginal effect of $x_{.j}$. This partial dependence function can also be extended to account for the joint marginal effect of two variables.

5 Extending the model to two-dimensional toxicity profiles

More general exposure escalation protocols involve the observation of a biological outcome y in association with a prescription of dose escalation $d \in [0, D]$, observed for a series of exposure times $t \in [t_0, T]$. Letting k, $(k = 1, \ldots, K)$ be a replication index, we define $y_{ik}(d, t)$ as the outcome of interest, evaluated at dose d, time t. The model in (1) can be extended as follows:

$$y_{ik}(d,t) \quad = \quad f(\boldsymbol{x}_i, d, t) + \varepsilon_{ik}(d,t), \tag{18}$$

where f is a random mean response surface and $\varepsilon_{ik}(d,t) \sim N(0, \sigma^2_{dt})$. In order to account for dependence between doses and durations of exposure, for each particle i, we assume

$$\mathrm{Cov}(\varepsilon_{ik}(d,t), \varepsilon_{ik'}(d',t')) = \sigma^2 \, \varphi_D^{|d-d'|} \, \varphi_T^{|t-t'|},$$

where $\varphi_D \in [0,1]$ and $\varphi_T \in [0,1]$ are autocorrelation parameters.

The response surface f in the terminal nodes of $\mathcal{T}$ is modeled by a tensor product of two one-dimensional P-splines (Lang and Brezger (2004)). Let $\mathcal{B}_1(.), ..., \mathcal{B}_{m_D+\delta}(.)$ be defined as in §2.2 and $\mathcal{B}_1(.), ..., \mathcal{B}_{m_T+\zeta}(.)$ denote $m_T + \zeta$ B-spline basis functions of order ζ on $[t_0, T]$, with m_T fixed knots. Then, if $\boldsymbol{x}_i$ is in the subset corresponding to the r^{th} terminal node of T_j,

$$f(\boldsymbol{x}_i, d, t) \quad = \quad \sum_{\ell=1}^{m_D+\delta} \sum_{m=1}^{m_T+\zeta} \beta_{r\ell m} \mathcal{B}_\ell(d) \mathcal{B}_m(t), \tag{19}$$

where $\boldsymbol{\beta}_r = (\beta_{r11}, ..., \beta_{r(m_D+\delta)(m_T+\zeta)})'$ is a vector of spline coefficients associated to the r^{th} terminal node.

The prior model has the same global dependence structure as in §3, but now includes an additional independent term φ_T for time-covariance. Let $t_0 = t_1 < ... < t_{n_T} = T$ be the design serie of exposure times. We adapt prior (9) to preserve conjugacy as follows:

$$p(\varphi_D) \quad \propto \quad \left(1 - \varphi_D^2\right)^{-\frac{n_T(n_D-1)}{2}} \exp\left(-\frac{\lambda_{01} - \varphi_D \lambda_{02} + \varphi_D^2 \lambda_{03}}{2\left(1-\varphi_D^2\right)}\right) \mathbb{I}\{\varphi_D \in [0,1]\}, \tag{20}$$

and introduce a similar distribution for φ_T:

$$p(\varphi_T) \quad \propto \quad \left(1 - \varphi_T^2\right)^{-\frac{n_D(n_T-1)}{2}} \exp\left(-\frac{\gamma_{01} - \varphi_T \gamma_{02} + \varphi_T^2 \gamma_{03}}{2\left(1-\varphi_T^2\right)}\right) \mathbb{I}\{\varphi_T \in [0,1]\}, \tag{21}$$

where γ_{01}, γ_{02} and γ_{03} are obtained by summing elements of the diagonal, subdiagonal and super-diagonal of matrix parameter prior Γ, constructed following the guidelines introduced in §3.3.

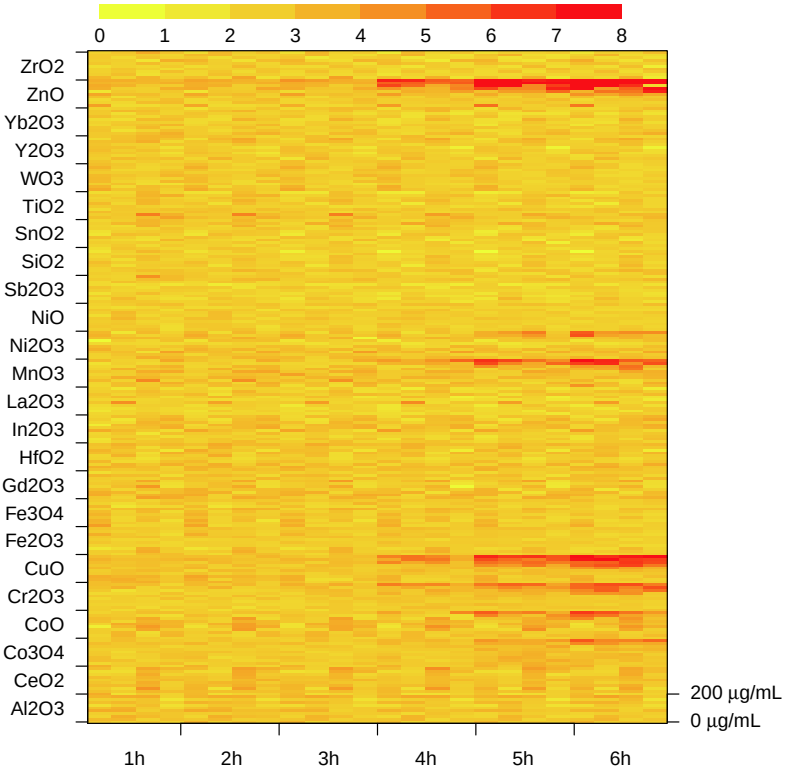

Figure 2: Heatmap for PI assay.

For terminal nodes splines coefficients priors, we use the spatial extension of Besag and Kooperberg (1995) of the first order random walk prior based on the four nearest neighbours of splines coefficients, with appropriate changes for corners and edges:

$$\boldsymbol{\beta}_r|\mathcal{T},\tau^2 \quad \propto \quad \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}_r' K_\beta \boldsymbol{\beta}_r\right), \tag{22}$$

where K_β is a penalty band matrix of size $(m_D+\delta)(m_T+\zeta)\times(m_D+\delta)(m_T+\zeta)$, which extends matrix (8) to the two-dimensional case. For posterior inference, we add a step to generate φ_T to the the Gibbs sampler of Section 4.

6 Applications

6.1 Case Studies Background

In this section, we illustrate our approach with experimental results from a case study reported by Zhang *and others* (2012), measuring toxicity of 24 metal oxides on human bronchial epithelial (BEAS-2B) cells. After 24h, Lactate Dehydrogenase (LDH) release was used to measure the death rate of cells exposed to eleven doses of metal oxides (from 0 to 200μg/mL), evenly spaced on the logarithmic scale. Cell death is commonly used to screen for ENM cytotoxicity without reference to a specific mechanism. Figure 1 shows the LDH dose-responses curves for the 24 metal oxide nanoparticles. Propidium Iodide (PI) fluorescence was used to indicate the percentage of cells experiencing oxidative stress through cellular surface membrane permeability, across the same ten doses and after six times of exposure (from 1 to 6h, at every hour). Figure 2 shows a heatmap representation for the PI assay, for all metal oxides, doses, times and replicates, where responses are color-coded from yellow (low) to red (high). In both assays, seven metal oxides (Co_3O_4, CoO, Cr_2O_3, CuO, Mn_2O_3, Ni_2O_3 and ZnO) display a notable rise for the higher doses, suggesting toxicity.

All metal oxides are characterized by six physio-chemical properties of potential interest to explain toxicity profiles: particle size in media, a measure of the crystalline structure (b(Å)), lattice energy ($\Delta H_{lattice}$), which measures the strength of the bonds in the particles, the enthalpy of formation ($\Delta H_{Me^{n+}}$), which is a combined measure of the energy required to convert a solid to a gas and the energy required to remove n electrons from that gas, metal dissolution rate, and conduction band energy, noted Ec (the energy to free electrons from binding with atoms).

In our analysis, we use cubic splines, i.e. $\delta = \zeta = 4$ and place interior knots at each intermediate dose from 0.39 to 100 μg/mL. Therefore, $n_D = m_D + 2$ and $n_T = m_T + 2$. For the tree prior, we adopt the default choice of Chipman et al. (2010), $(\alpha, \nu) = (0.95, 2)$, which puts more weight on trees of size 2 or 3. We place relatively diffuse priors $Gamma(1, 1)$ on precision parameters $1/\tau^2$ and $1/\sigma^2$. We choose $\Lambda = Id_{n_D}$ and $\Gamma = Id_{n_T}$ the identity matrices of size $n_D \times n_D$ and $n_T \times n_T$, assuming no prior correlations between measurements at different doses and times. Finally, moves "Grow", "Prune", "Change" and "Swap" of the Metropolis-Hastings tree-generating algorithm have probabilities 0.1, 0.1, 0.6 and 0.2, respectively. In our experiments, the high probability of the

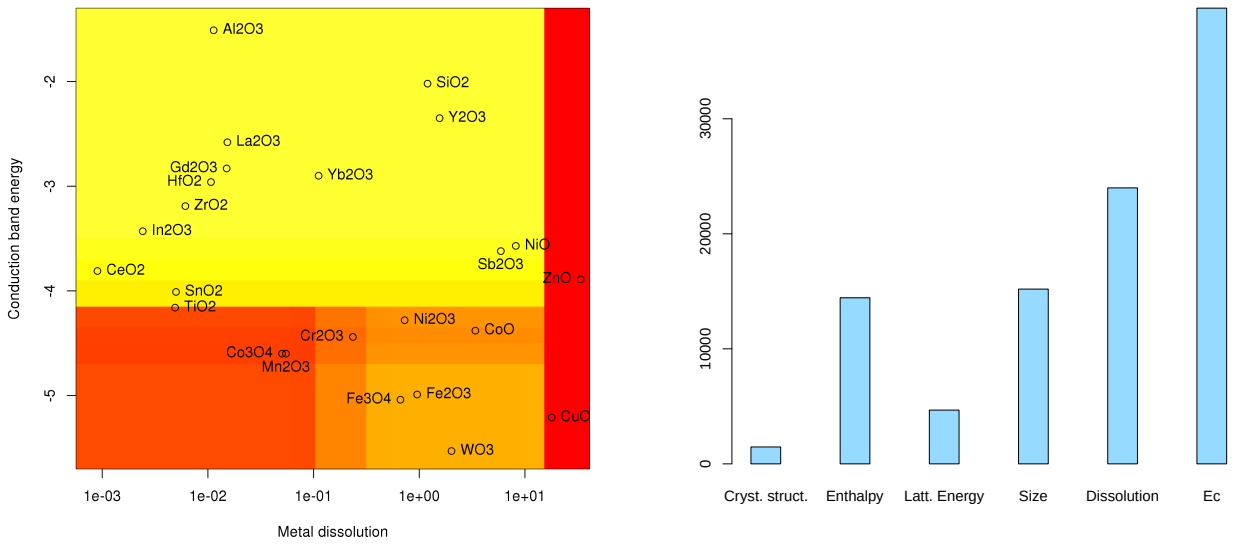

Figure 3: **LDH assay.** *(Left)* Marginal effect of metal dissolution (log scale) and conduction band energy on LDH. *(Right)* Variable importance scores for LDH assay.

"Change" move was sufficient to improve mixing and we did not add the complex move of Wu et al. (2007).

The MCMC algorithm of Section 4 was implemented in C++ and uses the Boost uBLAS library for efficient matrix manipulation (Walter and Koch (2010)). The rest of this section shows the results obtained on LDH and PI assays, based on 20000 iterations after discarding 80000 iterations for burn-in. Additional experimental results can be found in a supplement available from the authors.

6.2 LDH dose-escalation assay

We first perform posterior predictive checks for model fitting. Figure 4 shows the expected posterior predictive dose-response curves for two non-toxic metal oxides (CeO_2 and Fe_3O_4), and two toxic ones (Cr_2O_3 and ZnO), with the associated 90% intervals. All four intervals have good coverage of the original data. The other 20 curves exhibit similar behavior and can be found in a supplementary document.

Figure 3 *(Right)* shows the number of splits for the six physio-chemical properties over 20000

trees. High scores (over 20000, i.e. at least one split per tree on average) for metal dissolution and conduction band energy indicate that these two properties might play an important role in the definition of toxicity.

Figure 3 *(Left)* shows the combined marginal effect of conduction band energy and dissolution on LDH, color-coded from yellow (low) to red (high), for dose 200 μg/mL. The tree isolates a first region of high toxicity, corresponding to ENM with high dissolution rates (ZnO and CuO). This region corresponds to the first mechanism of toxicity identified by Zhang *and others* (2012): highly soluble metal oxides, such as ZnO and CuO, are more likely to release metal ions and disturb the cellular state.

A second region of toxicity on Figure 3 *(Left)* includes metal oxides Co_3O_4, CoO, Cr_2O_3, Mn_2O_3 and Ni_2O_3, with Ec values ranging from -4.33 eV for Mn_2O_3 to -4.59 eV for Ni_2O_3. This region matches the second mechanism for toxicity described by Zhang *and others* (2012): the overlap of the conduction band energy of the metal oxides with the biological redox potential of cells, ranging from -4.12 to -4.84 eV. When these two energy levels are alike, transfer of electrons from metal oxides to cells is facilitated, disturbing the intracellular state. Note that Figure 3 *(Left)* also shows an additional split that isolates Mn_2O_3, whose toxicity for LDH assay is more comparable to ZnO and CuO (see Figure 1). Similar figures for other doses are included in supplementary materials.

6.3 PI general exposure assay

Figure 6 illustrates the posterior predictive 90% surface intervals for two non-toxic metal oxides (La_2O_3 and TiO_2), and two toxic ones (Co_3O_4 and CuO), showing good posterior coverage over all doses and times of exposure. Similar surfaces for the other 20 metal oxides are plotted in supplementary materials.

Figure 5 *(Left)* shows the variable importance scores of the six physio-chemical properties over 20000 trees. Figure 5 *(Right)* illustrates the marginal effect of both conduction band energy and dissolution on membrane damage, color-coded from yellow to red, for dose 200 μg/mL and time $6h$. The tree model for PI assay also identifies the two areas of toxicity indicated in Zhang *and others* (2012), corresponding to highly soluble metal oxides and nanoparticles whose conduction band energy overlaps with cellular redox potential range. Additional figures for marginal effect of conduction band energy and metal dissolution, for all doses and times, are included in a supplementary

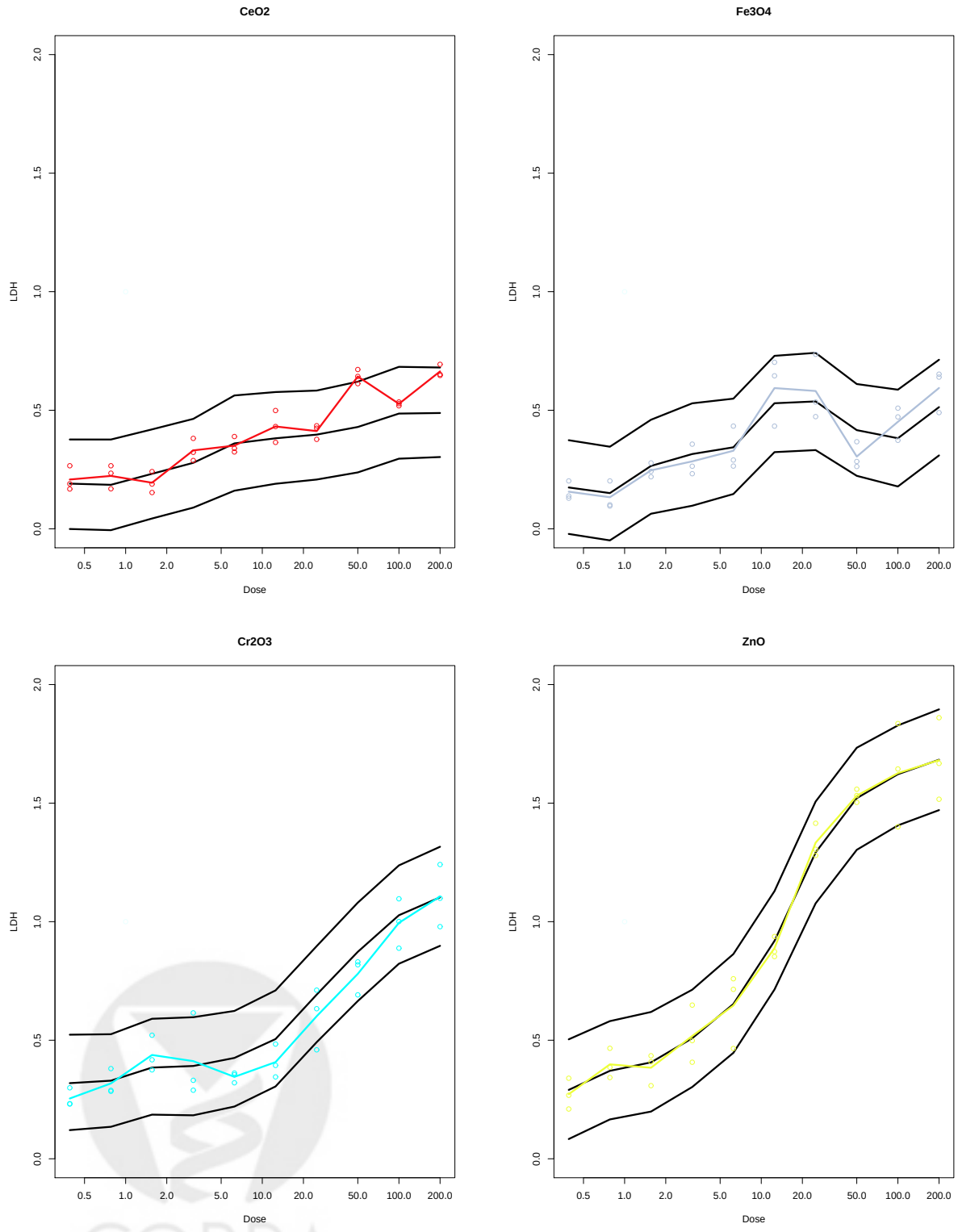

Figure 4: Posterior predictive curves for CeO_2, Fe_3O_4, Cr_2O_3 and ZnO. In color, the original data points and the average across replicates. In black, the expected posterior predictive curve and 90% interval.

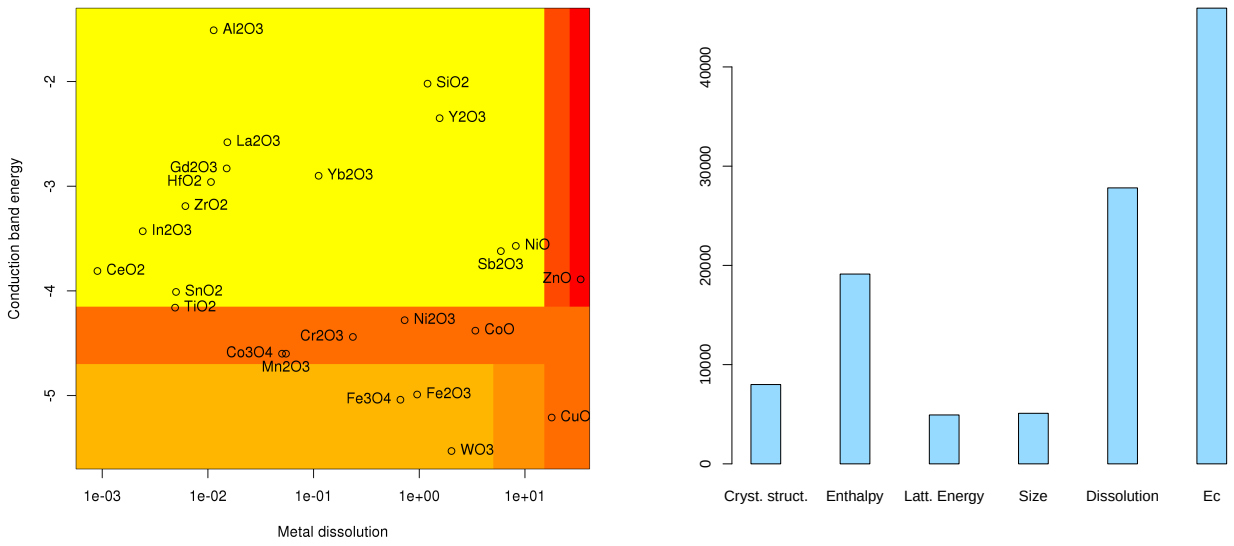

Figure 5: **PI uptake.** *(Left)* Marginal effect of metal dissolution (log scale) and conduction band energy on PI. *(Right)* Variable importance scores.

document.

The similarity of variable importance scores and marginal effect of conduction band energy and dissolution obtained for LDH and PI assays indicates a strong correlation between these assays for nanoparticle toxicity assessment, as noted by Zhang *and others* (2012).

7 Discussion

We propose a Bayesian regression tree model to define relationships between physio-chemical properties of engineered nanomaterials and their functional toxicity profiles in dose-escalation assays. As demonstrated by the case studies, the tree structure is adapted to account for flexible models of structure-activity relationships, such as threshold effects and interactions. The proposed model integrates information across all doses and replicates, and therefore is adapted to small sample sizes usually found in nanotoxicology datasets. The posterior distribution integrates over the model space to provide straightforward inference on non-trivial functionals of parameters of interest. The smoothing splines representation allows for easy extension of the model to two-dimensional toxicity profiles of general exposure escalation assays as well as for modeling sub-lethal outcomes.

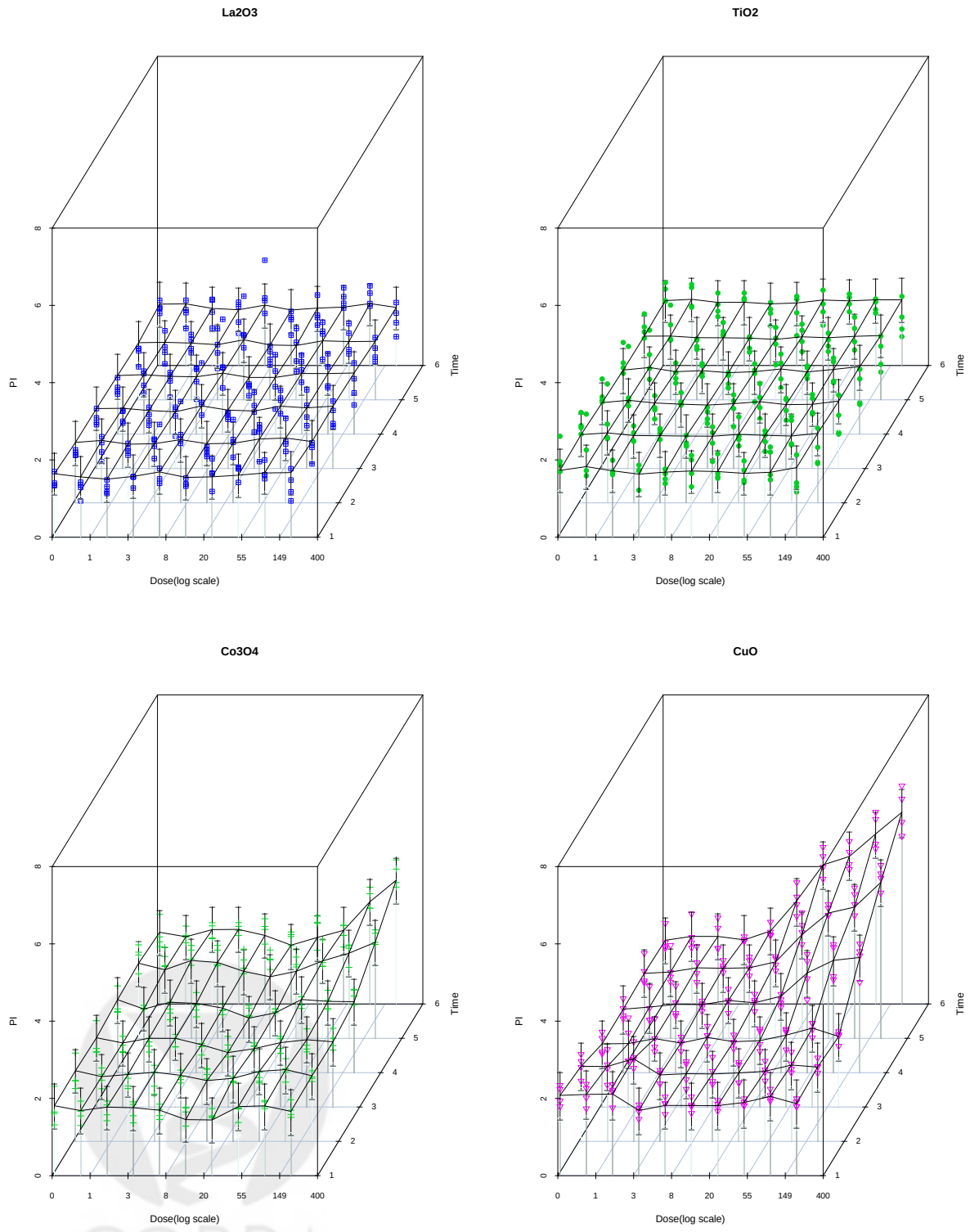

Figure 6: Posterior predictive curves for La_2O_3, TiO_2, Co_3O_4 and CuO. In color, the original data points. In black, the expected posterior predictive surface and 90% interval.

One potential pitfall of the method is the use of a single tree model, instead of ensemble methods such as random forests (Breiman (2001)) or sum-of-trees models (Chipman et al. (2010)). Sum-of-trees models are known to mix better and to incorpore more easily interactions and additive effects (Chipman et al. (2010)). However, this could result in very heavy computations, as each tree has to handle multidimensional structure in nodes and large penalization matrices. Another drawback of the proposed methodology is the normality assumption on the measurement error, which does not account for potential heterogeneity of responses across nanoparticles. This assumption can be relaxed by introducing an nanoparticle-specific random variance inflation parameter, that results marginally in a t-distribution for the errors (Patel et al. (2012)).

The main goal of this work is to provide a statistically sound framework for predictive nanotoxicology. Typical QSAR models often rely on unwarranted or arbitrary data summaries, with hardly assessed consequences on predictive reliability and inferential characterization of uncertainty. The Bayesian framework, averaging over models and smoothing structures, is likely to provide a more realistic account of structural uncertainty. Alternative inferential frameworks would require to split the dataset in a learning set and testing set for external validation. However, the small sample size and unbalanced structure of current nanotox data, could affect the predictive performance of the model and require adapted resampling techniques, such as under-sampling or over-sampling. An investigation aimed at comparing different inferential paradigms is beyond the scope of this manuscript, but it is certainly very important and worth pursuing. For an example related to linear model determination, see the paper by Celeux et al. (2012).

As seen in the case study, different toxicity mechanisms can be closely related. Therefore, an important opportunity for model extensions would be to combine different biological assays in a single analysis through multiple trees. The final goal being that of understanding if particles physical and chemical properties have a differential effect on different cellular injury pathways. These potential developments are indeed much needed. However, more sophisticated modeling strategies will be more likely to be useful if technological advances will allow for feasible screening of much larger nanomaterial libraries.

8 Supplementary Material

Supplementary material and C++ code to perform the experiments and the simulation dataset are available upon request to the author.

Acknowledgements

Primary support was provided by the U.S. Public Health Service Grant U19 ES019528 (UCLA Center for Nanobiology and Predictive Toxicology). This work was also supported by the National Science Foundation and the Environmental Protection Agency under Cooperative Agreement Number DBI-0830117. Any opinions, findings, conclusions or recommendations expressed herein are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Environmental Protection Agency. This work has not been subjected to an EPA peer and policy review.

References

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82,** 733–746.

Breiman, L. (2001). Random Forests. *Machine Learning* **45,** 5–32.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Celeux, G., El Anbari, M., Marin, J.-M., and Robert, C. P. (2012). Regularization in regression: comparing Bayesian and Frequentist methods in a poorly informative situation. *Bayesian Analysis* **7,** 477–502.

Chipman, H., George, E. I., and McCulloch, R. E. (2002). Bayesian treed models. *Machine Learning* **48,** 299–320.

Chipman, H. A., George, E. I., and Mcculloch, R. E. (1997). Bayesian CART model search. *Journal of the American Statistical Association* **93,** 935–960.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics* **4,** 266–298.

De'ath, G. (2002). Multivariate regression trees: A new technique for modeling species-environment relationships. *Ecology* **83,** 1105–1117.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 89–121.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* **29,** 1189–1232.

Galimberti, G. and Montanari, A. (2002). *Regression trees for longitudinal data with time-dependent covariates*, pages 391–398. Classification, Clustering and Data Analysis. Springer, New York, k. jajuga, a. sokolowski, and h.-h. bock edition.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13,** 183–212.

Liu, R., Rallo, R., George, S., Ji, Z., Nair, S., Nel, A. E., and Cohen, Y. (2011). Classification nanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small* **7,** 1118–1126.

Müller, P., Shih, Y.-C. T., and Zhang, S. (2007). A spatially-adjusted Bayesian additive regression tree model to merge two datasets. *Bayesian Analysis* **2,** 611–633.

Pang, O. (2009). *On the implementation and extension of BART.* PhD thesis, University of Pennsylvania.

Patel, T., Telesca, D., George, S., and Nel, A. E. (2012). Toxicity profiling of engineered nanomaterials via multivariate dose-response surface modeling. *Annals of Applied Statistics* **6,** 1707–1729.

Rowe, D. B. (2003). *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing.* Chapman & Hall/CRC, Boca Raton, FL.

Segal, M. and Xiao, Y. (2011). Multivariate random forests. *WIREs Data Mining Knowledge Discovery* **1,** 80–87.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association* **87,** 407–418.

Sela, R. J. and Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning* **86,** 169–207.

Walter, J. and Koch, M. (2010). The uBLAS library.

Wu, Y., Tjelmeland, H., and West, M. (2007). Bayesian CART: Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics* **16,** 44–66.

Yu, K., Wheeler, W., Li, Q., Bergen, A. W., Caporaso, N., Chatterjee, N., and Chen, J. (2010). A partially linear tree-based regression model for multivariate outcomes. *Biometrics* **66,** 89–96.

Yu, Y. and Lambert, D. (1999). Fitting trees to functional data, with an application to time-of-day patterns. *Journal of Computational and Graphical Statistics* **8,** 749–762.

Zhang, H., Ji, Z., Xia, T., Meng, H., Low-Kam, C., Liu, R., Pokhrel, S., Lin, S., Wang, X., Liao, Y.-P., Wang, M., Li, L., Rallo, R., Damoiseaux, R., Telesca, D., Mädler, L., Cohen, Y., Zink, J. I., and Nel, A. E. (2012). Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation. *ACS Nano* **6,** 4349–4368.