Bayesian Matrix Completion for Hypothesis Testing

Bora Jin[1,*], David B. Dunson[1], Julia E. Rager[2], David M. Reif[3], Stephanie M. Engel[2], and Amy H. Herring[1]

[1] *Duke University, Durham, USA.*

[2] *University of North Carolina at Chapel Hill, Chapel Hill, USA.*

[3] *North Carolina State University, Raleigh, USA.*

E-mail: *bora.jin@duke.edu

Summary. High-throughput screening (HTS) is a well-established technology that rapidly and efficiently screens thousands of chemicals for potential toxicity. Massive testing using HTS primarily aims to differentiate active vs inactive chemicals for different types of biological endpoints. However, even using high-throughput technology, it is not feasible to test all possible combinations of chemicals and assay endpoints, resulting in a majority of missing combinations. Our goal is to derive posterior probabilities of activity for each chemical by assay endpoint combination, addressing the sparsity of HTS data. We propose a Bayesian hierarchical framework, which borrows information across different chemicals and assay endpoints in a low-dimensional latent space. This framework facilitates out-of-sample prediction of bioactivity potential for new chemicals not yet tested. Furthermore, this paper makes a novel attempt in toxicology to simultaneously model heteroscedastic errors as well as a nonparametric mean function. It leads to a broader definition of activity whose need has been suggested by toxicologists. Simulation studies demonstrate that our approach shows superior performance with more realistic inferences on activity than current standard methods. Application to an HTS data set identifies chemicals that are most likely active for two disease outcomes: neurodevelopmental disorders and obesity. Code is available on Github.

Keywords: Bayesian hierarchical model; bioactivity profiles; dose-response modelling; heteroscedasticity; latent factor models; high-throughput screening; multiple testing.

1. Introduction

Screening and regulating hazardous chemicals is of great importance and urgency especially as massive numbers of new chemicals are introduced (an average of 2,000 chemicals per year). The traditional animal or *in vivo* testing paradigms are infeasible due to financial and time constraints (Dix et al., 2007; Judson et al., 2010); in addition, it is desirable to minimise animals used in any testing procedure for ethical reasons. Many organisations such as the World Health Organization's Intergovernmental Forum on Chemical Safety (WHO - IFCS), European Chemicals Agency (ECHA), and United States Environmental Protection Agency (EPA) screen chemicals to measure their potential toxicity and develop alternatives to animal testing.

As a high-throughput screening (HTS) mechanism has been developed based on *in vitro* assays and a large number of chemicals, EPA and ECHA have had opportunities to operate relatively low-cost and rapid chemical screening programs. For instance, Toxicity Forecaster (ToxCast) and Toxicology in the 21st Century (Tox21) from EPA are designed to identify chemicals that likely induce toxicity in humans and prioritise them for further testing (Judson et al., 2010), using HTS methods. ECHA also promotes similar approaches expediting chemical risk prioritisation and assessment for the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation (ECHA, 2017). For the rest of this paper, we use the EPA's ToxCast/Tox21 data as a representative of general HTS data. To the authors' knowledge, ToxCast/Tox21 is the largest in size among existing HTS data in toxicology. Furthermore, it has universal coverage of molecules that many regions including Canada, Japan, and European Union as well as the United States have approved for clinical use (Tice et al., 2013).

The ToxCast/Tox21 program tests thousands of chemicals against numerous high-throughput assay endpoints. Although the HTS mechanism has provided a relatively cheap and quick way to conduct millions of tests, it is still only possible to test a small minority of all (chemical, assay endpoint) combinations. This leads to many non-tested combinations, which are shown as empty cells in Figure 1 and Figure S1 in the Supporting Material. Figure 1 displays a few cells from the ToxCast/Tox21 data, including the observed measurements. Some cells have few observations, while others have multiple replicates at a finer grid of doses. Here, an observation is the result from an experiment where a chemical is applied to an assay endpoint at a certain dose. Figure S1 illustrates the overall structure of the data, colour-coded by the number of observations; many cells have zero observations, and the number of observations largely fluctuates across cells. Such HTS data can be arranged as matrix-structured sparse functional data, with the rows of the matrix corresponding to different chemicals, the columns to different assay endpoints, and the cells containing dose-response measurements.

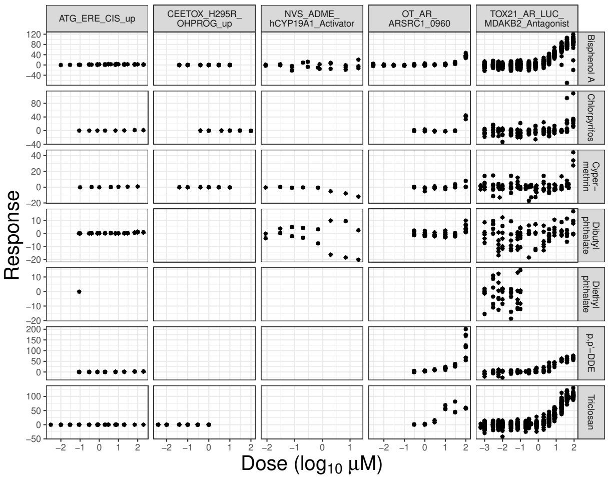

Fig. 1. Detailed illustration of the ToxCast/Tox21 data structure. Sample data for 7 chemicals (rows) and 5 assay endpoints (columns). Each cell contains a set of test results of a single chemical against one assay endpoint, which is functional data on a dose-response curve.

Traditional matrix completion focuses on the problem of filling in the missing elements of a large matrix based on observations on a small proportion of the cells. Typically, the observed cells contain a scalar that is assumed to be measured without error. On the other hand, we are faced with a latent binary matrix completion problem with each cell containing a binary indicator of whether a particular chemical is active for a specific assay endpoint. This can be viewed as a matrix-structured multiple hypothesis testing problem for assessing dose-response relationships.

There have been many Bayesian approaches to multiple hypothesis testing (Scott and Berger, 2006; Thomas et al., 2009; Scott and Berger, 2010; Li and Zhang, 2010; Scheel et al., 2013; Wilson et al., 2014). Bayesian approaches are attractive due to their automatic adjustment for multiplicity (Scott and Berger, 2006, 2010) by treating hyperparameters controlling model size as unknown and informed by the data. In the typical framework, hypotheses are considered exchangeable *a priori*. For example, variable selection cases have hypotheses $H_{0j} : \gamma_j = 0$ and $H_{1j} : \gamma_j = 1$ in which γ_j is an indicator of whether the jth variable is included for $j = 1, \ldots, p$, and $\pi_0 = Pr(\gamma_j = 1) \sim Beta(a, b)$ is a global parameter controlling model size. A variety of more elaborate non-exchangeable priors have been proposed for $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, designed to include "prior covariates" Z_j informing $Pr(\gamma_j = 1)$ (Thomas et al., 2009) and known structure among covariates represented by an undirected graph (Li and Zhang, 2010).

There has also been some consideration of matrix-structured multiple testing. In relation to dose-response curves, Wilson et al. (2014) test for dose effects on the mean using a generalised linear mixed effects model. The mean effect indicator γ_{ij} for a (chemical i, assay endpoint j) pair follows a Bernoulli distribution with $\pi_{ij} = Pr(\gamma_{ij} = 1)$. Then π_{ij} is further structured with an assay endpoint random effect, chemical-level fixed effect and a probit link: $\pi_{ij} = \Phi(\alpha_j + \alpha x_i)$ where x is the chemical-level covariate. However, it is nontrivial to find informative chemical-level covariates x_i in this context. In their ToxCast/Tox21 application, Wilson et al. (2014) found their covariate, chemical solubility, was not significant in explaining γ_{ij}, resulting in a simplified model with only random effects for assay endpoints.

In order to account for mechanistic similarities among chemicals and/or assay endpoints as well as to tackle sparsity of the data, we require a more sophisticated hierarchy that borrows information across both rows and columns of the matrix. Tansey et al. (2019) propose hierarchical functional matrix factorisation methods to infer dose-response curves, approximating the row and the column space using low-dimensional latent attributes. However, their model lacks a formal testing framework. Furthermore, they assume a matrix data structure in which all cells have the same number of replicates at the same number of unique doses, which might not be guaranteed in HTS data. To illustrate, the ToxCast/Tox21 data have different numbers of unique doses within a column and varying numbers of replicates at each dose within a cell.

Thus, we adapt low rank approximations addressing matrix completion problems (Mnih and Salakhutdinov, 2008; Koren et al., 2009; Purushotham et al., 2012; Tansey et al., 2019) to a multiple hypothesis testing framework and extend them for more general data structures. This hierarchical Bayesian matrix completion (BMC) approach for hypothesis testing is particularly useful for sparse data. We construct π_{ij} with a latent factor model, assuming that low-dimensional latent attributes account for associations relevant to the mean effect among chemicals or assay endpoints. A posterior summary matrix of γ_{ij} naturally prioritises chemicals and enables out-of-sample prediction of bioactivity for new chemicals not yet assayed.

Other important characteristics of HTS data are irregular dose-response shapes and heteroscedasticity. Many previous studies placed monotone non-decreasing shape restrictions on dose-response curves (Neelon and Dunson, 2004; Ritz, 2010; Wilson et al., 2014) and did not consider heteroscedasticity. Our approach is strongly motivated by evidence that disruption in centrality or dispersion of intricately-controlled biological pathways observed *in vitro* can lead to *in vivo* toxicity and ultimately connect to detrimental health effects (Klaren et al., 2019; Knapen et al., 2020). Accordingly, a novel attempt in toxicology simultaneously to model heteroscedastic errors as well as any non-constant shapes of the mean completes BMC. This leads to a broader definition of activity as any changes in mean and variance of dose-response curves. These considerations provide a more holistic perspective on active chemicals than previous research.

The remainder of the paper is organised as follows. Section 2 further explains motivating aspects for a model that is applicable to HTS data. Section 3 summarises the ToxCast/Tox21 data of relevance to neurodevelopmental disorders and obesity. Then, the BMC approach is described by subparts throughout section 4. We compare the performance of BMC with existing methods on simulated data sets and show results using our HTS data in section 5, highlighting chemicals that pose greater risks for obesity and neurodevelopment endpoints. Potential areas of future research are discussed in section 6.

2. Motivating Aspects and Relevant Literature

2.1. Hierarchical Structures

A simple approach for matrix-structured data would be to consider each cell independently. The EPA has developed an R package "tcpl" (Filer et al., 2017) to facilitate independent dose-response modelling of the ToxCast data. This R package provides three default models: a constant model at zero, a three-parameter Hill model, and a five-parameter gain-loss model for each (chemical, assay endpoint) combination separately. Unfortunately, independently inferring dose-response relationships does not have predictive power: it cannot predict activity for cells having no data. Further, it is likely to have low power and high variance in estimation due to the intrinsic sparsity of the ToxCast/Tox21 data shown in Figures 1 & S1. In the ToxCast/Tox21 data, the median number of unique doses tested for each pair is 8 (Figure S2 in the Supporting Material), and about 30% of them are without replicates. Therefore, hierarchical methods for borrowing information are crucial.

2.2. Splines without Shape Restrictions

In estimating dose-response curves, researchers have often forced parametric or monotone restrictions on shapes of the curves to increase interpretability. The EPA's default models currently available through the tcpl package heavily depend on parametric assumptions and are restricted to positive responses to reduce the parameter space, requiring an inverse transformation to fit negative responses. In addition, Wilson et al. (2014) model dose-response functions by piecewise log-linear splines with constrained parameters to ensure responses are monotone and non-decreasing. In the ToxCast/Tox21 data, it appears difficult to standardise shapes of the dose-response curves (Figure 1). Furthermore, we observe some examples of decreasing trends between certain assay endpoints (e.g., TOX21_ERa_LUC_BG1_Agonist as shown in Figure S3 in the Supporting Material) and multiple chemicals. Thus, we propose a non-restricted spline model robust to any shapes of dose-response curves, given that both upturns and downturns in dose-response functions are suggestive of potential toxicity.

2.3. Heteroscedastic Variances

Toxicological HTS data have innate heteroscedasticity. Such heteroscedasticity is inevitable because dose effects are variable by nature, with variability often amplified at high doses. Differences in the ability of assays to absorb chemical doses further inflate this variability. Wilson et al. (2014) attempted to reduce such heteroscedasticity by log transforming the data. However, data transformations may be hard to justify theoretically (Leslie et al., 2007) and may be insufficient practically. In genetics, multiple studies have been conducted to detect genetic loci that affect heteroscedastic errors of quantitative traits of interest (Paré et al., 2010; Rönnegård and Valdar, 2012; Yang et al., 2012). It is widely appreciated that analysing differences in variance could reveal a previously unknown genetic influence and alternative biological relevance. Although detection of heteroscedastic variances is routinely considered in genetic analysis (Corty and Valdar, 2018), it has not been of main interest in chemical toxicity analysis. Without data transformations, we consider heteroscedasticity as another source of information. We compute posterior probabilities of the variance effect for observed cells.

3. Data

This paper uses data from the ToxCast/Tox21 project (invitroDBv3.2, released on March 2019), available at `https://epa.figshare.com/articles/ToxCast/Tox21_Database_invitroDB_for_Mac_Users/6062620`. We focus on a subset of the ToxCast/Tox21 data that contain assay endpoints relevant to neurodevelopmental disorders and obesity, along with chemicals tested over those assay endpoints. As a result of selection criteria for chemicals and assay endpoints described in the Supporting Material S2, 30 chemicals evaluated across 131 assay endpoints are studied for neurodevelopmental disorders. These create in total 3930 cells, from which 2024 cells (51.5%) are missing. For obesity, we use the same 30 chemicals evaluated across 271 assay endpoints. Among the total 8130 cells, 3274 cells (40.3%) contain no data.

4. Model

We specify motivations for each part of the model including the prior specification.

4.1. Matrix Completion

Primary interest lies in differentiating active and inactive chemicals. First, we conduct multiple hypothesis testing of whether the dose-response curve is constant or not across chemicals and assay endpoints. We introduce latent binary indicators $\{\gamma_{ij}\}$, with $\gamma_{ij} = 1$ denoting that the average dose-response curve is not constant for the (chemical i, assay endpoint j) pair. Let vector $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iJ})^T$ represent chemical i's *mean effect profile* across the J assay endpoints for $i = 1, \ldots, m$. We assume that chemicals and assay endpoints explain dose effects on the mean via low rank latent features, for which we exploit a sparse Bayesian factor model (Bhattacharya and Dunson, 2011). Since each γ_{ij} takes $\{0, 1\}$ values, we impose a generalised factor model using a probit link:

$$Pr(\gamma_{ij} = 1) = \pi_{ij} = \Phi(\boldsymbol{\lambda}_i^T \boldsymbol{\eta}_j). \tag{1}$$

A data-augmented form rewrites (1) as

$$\gamma_{ij} = \mathbf{1}(z_{ij} > 0) \text{ where } z_{ij} \sim N(\boldsymbol{\lambda}_i^T \boldsymbol{\eta}_j, 1). \tag{2}$$

In this factor model, λ_{il} represents the coefficient of the lth latent pathway for the ith chemical to have the mean effect, and η_{lj} represents that for the jth assay endpoint to have the mean effect for $l = 1, \ldots, q$ and $q \ll \min(m, J)$. The inequality is reasonable, assuming that not every assay endpoint (or chemical) forms an idiosyncratic latent pathway for the mean effect. The ToxCast/Tox21 application lets either $\boldsymbol{\lambda}_i$ be treated as factor loadings and $\boldsymbol{\eta}_j$ as latent factors or vice versa, depending on researchers' interests. Provided that one is interested in latent covariance structure among chemicals with regards to the mean effect, a standard factor model puts a multivariate standard normal prior on latent factors $\boldsymbol{\eta}_j \sim N_q(\mathbf{0}, I)$. Integrating out $\boldsymbol{\eta}_j$ from (2) yields $\mathbf{z}_j \sim N_m(\mathbf{0}, \Lambda\Lambda^T + I)$ where Λ has $\boldsymbol{\lambda}_i^T$ as its ith row. This factor model provides a low dimensional representation of the underlying covariance structure of chemicals. We employ a multiplicative gamma process shrinkage prior on factor loadings as in Bhattacharya and Dunson (2011):

$$\lambda_{il} \sim N(0, \phi_{il}^{-1}\tau_l^{-1}),\ \phi_{il} \sim Gamma\left(\nu/2, \nu/2\right),\ \tau_l = \prod_{h=1}^{l} \zeta_h,\ l = 1, \ldots, q, \tag{3}$$

$$\zeta_1 \sim Gamma(a_1, 1),\ \zeta_h \sim Gamma(a_2, 1),\ h \geq 2. \tag{4}$$

This prior choice is supported by Judson et al. (2010) who elucidate relationships between chemicals and published pathways. The authors discovered that chemicals are activating various human genes and pathways, but the number of activated pathways varies widely across chemicals. The multiplicative gamma process shrinkage prior tends to shrink columns of a loading matrix towards zero through the τ_l's. At the same time, it is possible to strongly shrink only a subset of elements in a certain column through local shrinkage parameters ϕ_{il}'s, retaining sparse signals.

As alternatives to the above model, we considered adding another parameter controlling the overall proportion of γ_{ij}'s equal to one. This can be accomplished through adding an intercept and letting $Pr(\gamma_{ij} = 1) = \Phi(\xi + \boldsymbol{\lambda}_i^T \boldsymbol{\eta}_j)$ or including an unknown mean at the latent variable level as $\lambda_{il} \sim N(\xi, \phi_{il}^{-1}\tau_l^{-1})$. We found that, in both these cases, adding an unknown ξ parameter did not improve results, and indeed can lead to worse performance. This is likely due to the fact that the extra parameter is effectively redundant, leading to an over-parametrised model. Hence, we set $\xi = 0$ as in (1) in all the analyses we report in this paper.

4.2. Dose-Response Functional Data Analysis

4.2.1. Splines without Shape Restrictions

Let x_{ijk} be a test dose (in log base 10 scale in micromolar (μM)) of the kth measurement for a (chemical i, assay endpoint j) pair, and let y_{ijk} be the corresponding response. Consider the model $y_{ijk} = \gamma_{ij} f_{ij}(x_{ijk}) + \epsilon^*_{ijk}$, where the error distribution is $\epsilon^*_{ijk} \sim N(0, \sigma^{*2}_{ijk})$ for $i = 1, \dots, m$, $j = 1, \dots, J$, and $k = 1, \dots, K_{ij}$. Non-constant dose-response curves are estimated when $\gamma_{ij} = 1$. We model the dose-response function f_{ij} using cubic B-splines, which is equivalent to estimating $\boldsymbol{\beta}_{ij}$ in $(f_{ij}(x_{ij1}), \dots, f_{ij}(x_{ijK_{ij}}))^T = X_{ij}\boldsymbol{\beta}_{ij}$ with the B-spline basis matrix X_{ij} of size $(K_{ij} \times p)$. We normalise responses and centre columns of the B-spline basis matrix by (i, j) pairs prior to any analyses in order to exclude the intercept. The prior distributions of spline coefficients and their hyperparameters are $\boldsymbol{\beta}_{ij} \overset{ind.}{\sim} N_p\left(0, \Sigma_j\right)$; $\Sigma_j^{-1} \overset{iid}{\sim} Wish_p(a, R^{-1})$ with fixed a and R, where $\Omega \sim Wish_p(m, A)$ is a Wishart distribution in p-dimensions with $E(\Omega) = mA$. We suggest the following default choices for our application. As Figure 1 suggests, dose-response functions share more similarities within an assay endpoint than between different assay endpoints. This suggests a formulation in which spline coefficients of different chemicals have a common prior covariance matrix for the same assay endpoint. For assay endpoint-specific covariance matrices, R is determined as the empirical covariance of the ordinary least squares estimates for chemical-assay endpoint pairs. The degrees of freedom parameter a is chosen to be $p+2$ so that Σ_j is loosely centred around R.

4.2.2. Heteroscedastic Variances

Figure S4 in the Supporting Material illustrates that ranges of responses may vary substantially by assay endpoints. This suggests modelling errors with assay endpoint-specific variances. Moreover, we are motivated to capture heteroscedasticity to explain another dimension of chemical activity. We use a log-linear model on σ^{2*}_{ijk} so that $\log \sigma^{*2}_{ijk} = \delta_{0j} + x_{ijk}\delta_{ij}$ and $\sigma^{*}_{ijk} = \exp(\delta_{0j}/2)\exp(x_{ijk}\delta_{ij}/2)$. Here, we separate variance into an assay endpoint-specific variability and a part that changes with dose. Reparametrising $\exp(\delta_{0j}/2)$ with σ_j gives the final model equation

$$y_{ijk} = \gamma_{ij} f_{ij}(x_{ijk}) + \exp(x_{ijk}\delta_{ij}/2)\epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0, \sigma_j^2). \tag{5}$$

The assay endpoint-specific variances have an inverse-Gamma distribution *a priori*: $1/\sigma_j^2 \overset{iid}{\sim} Gamma\left(\nu_0/2, \nu_0\sigma_0^2/2\right)$, ν_0, σ_0^2 fixed. In our application, we suggest fixing the hyperparameters ν_0 at 1 and σ_0^2 at the sample variance of the response variable to have the prior distribution weakly centred around a simple estimate from data. With respect to the heteroscedastic noise, let $t_{ij} = \mathbf{1}(\delta_{ij} \neq 0)$ so that t indicates activity causing *variance changes* of responses. Prior and hyperprior distributions of heteroscedasticity parameters are given similarly to Leslie et al. (2007): $t_{ij} \overset{iid}{\sim} Bernoulli(\pi_t)$; $\pi_t \sim Unif(0,1)$; $\delta_{ij}|t_{ij} = 1 \overset{iid}{\sim} N(0, v_\delta)$, $Pr(\delta_{ij} = 0|t_{ij} = 0) = 1$, v_δ fixed. In our case, we found that ensuring a large enough value for v_δ that appropriately covers the data range improves estimation of δ_{ij} and t_{ij}. Provided that conditional standard deviations of responses given doses can be proxies for $\exp(x_{ijk}\delta_{ij}/2)$ in (5), a range of δ_{ij}'s is obtained. The variance parameter v_δ of δ_{ij} is then determined as the square of the range divided by 4, which makes ±2-standard deviation intervals for δ_{ij} cover its sample range. We finally fix v_δ at the maximum of the above value and the sample variance of the response variable. Combined with σ_0^2, this allows the prior distributions of two variance parts - the assay endpoint-specific and the heteroscedastic variance - to place enough probability on the observed variability from the data.

In conclusion, (5) is the final model in which γ_{ij} is an indicator specifying whether the ith chemical activates the jth assay endpoint in the mean, f_{ij} is a dose-response function, the exponential term allows for heteroscedastic noise, and error is modeled with normal distributions having assay endpoint-specific variances.

4.3. Posterior Computation

Our posterior samples are obtained using Metropolis-Hastings steps within a partially collapsed Gibbs sampler. Most of the parameters have conjugate posterior distributions which lead to a straightforward Gibbs sampler. Details are provided in the Supporting Material S4 with code at `https://github.com/jinbora0720/BMC`.

5. Results

5.1. Simulations

Simulation studies were conducted to evaluate the performance of BMC in learning the latent correlation structure among chemicals, predicting the mean effect probabilities, and estimating the parameters. Two broad scenarios of simulations were examined corresponding to data simulated from BMC (Simulation 1) or an alternative (Simulation 2). For predictive performance, BMC was compared to three variations in the prior structure of γ_{ij}. Instead of a latent factor model, we assume simpler structures *a priori* as follows:

$$Pr(\gamma_{ij} = 1) = \pi_0 \ \forall i, j, \text{ and } \pi_0 \sim Beta(1, 1); \tag{6}$$

$$Pr(\gamma_{ij} = 1) = \pi_i \ \forall j, \text{ and } \pi_i \overset{iid}{\sim} Beta(1, 1) \ \forall i; \tag{7}$$

$$Pr(\gamma_{ij} = 1) = \pi_j \ \forall i, \text{ and } \pi_j \overset{iid}{\sim} Beta(1, 1) \ \forall j. \tag{8}$$

We call models with (6), (7), (8) BMC$_0$, BMC$_i$, and BMC$_j$, respectively. BMC$_i$ assumes that each chemical has its own intrinsic mean effect probability, while BMC$_j$ assumes that each assay endpoint has its own mean effect probability. For estimation performance, the proposed model is compared to the zero-inflated piecewise log-logistic model (ZIPLL) (Wilson et al., 2014) and tcpl (Filer et al., 2017). The ZIPLL code at `https://github.com/AnderWilson/ZIPLL` utilises a Bayesian hierarchical approach whose testing framework for the mean effect adopts (6). Since the code does not allow missing pairs in the data, we only use ZIPLL for estimation and not prediction. The tcpl models are currently used by EPA and treat dose-response curves independently.

In Simulation 1 in which BMC is the true data generating process, mimicking the ToxCast/Tox21 application, the number of chemicals m was set to 30, and the number of assay endpoints J to 150. We generate 50 data sets, and in each set we hold out 135 pairs at random, which are 3% of the total cells in the data matrix. The profiles of the mean effect for chemical-assay endpoint pairs were sampled assuming a factor model, which induced a correlation structure among chemicals (Figure S5 in the Supporting Material). For pairs having dose effects on the mean, dose-response functions were given as one of the three categories: mostly increasing and decreasing at higher doses; monotonically increasing; and decreasing. Figure 2 presents examples of dose-response functions of each category. Heteroscedasticity is expected at one fifth of chemical-assay endpoint combinations. More specific settings of Simulation 1 are described in the Supporting Material S3.

As illustrated in Figure 2, BMC accurately captures true curves regardless of shapes. It also produces tighter 95% credible intervals for the average dose-response curves than competitors. The competitors, ZIPLL and tcpl models, do not seem robust enough to various dose-response curves. In particular, ZIPLL estimates a decreasing trend as constant, which is evident in Figure 2C. For generally increasing curves (A, B), the ZIPLL and tcpl models sometimes miss the true dose-response functions, which becomes more noticeable when heteroscedasticity exists. These results suggest that in some cases, BMC can lead to more precise inferences on values estimated through dose-response curves, such as Emax (greatest attainable response) or AC50 (chemical dose producing half maximal response in an assay endpoint).

BMC also provides precise estimation of the latent correlations relevant to the mean

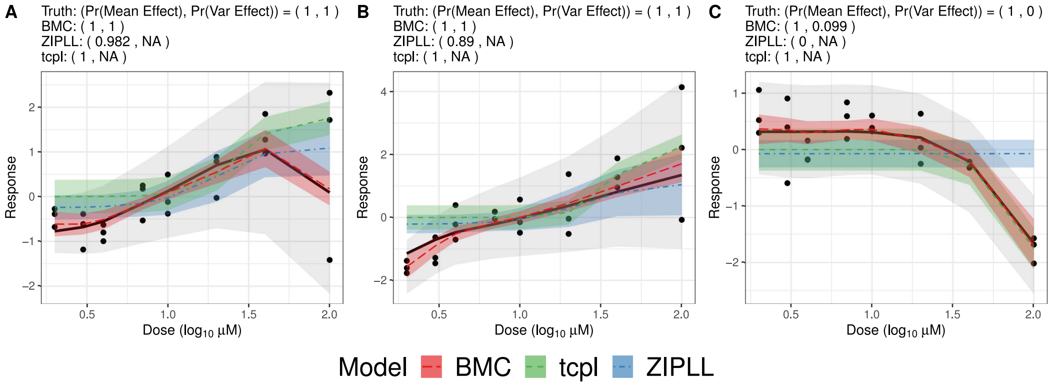

Fig. 2. Dose-response curves (solid line) with fitted mean functions by BMC (long dash line), ZIPLL (dot-dashed line), and tcpl (dashed line) in Simulation 1. The true curve is mostly increasing and decreasing at higher dose in A, monotonically increasing in B, and monotonically decreasing in C. Darker gray areas around estimated functions represent 95% credible intervals for the average dose-response curves computed by BMC and ZIPLL, and confidence intervals by tcpl. Lighter gray areas illustrate 95% predictive intervals from BMC for data points.

effect among chemicals (Figure S5 in the Supporting Material). Two factors ($q = 2$) generated the truth, and the sampler ran with a guess of three more factors. The multiplicative gamma process shrinkage prior helped recover the true number of factors $q = 2$ by shrinking factor loadings of redundant factors to zero (Figure S6 in the Supporting Material). Figure 3 displays an example of mean effect profiles. The truth is adequately captured via the estimated and predicted probabilities. Results from a 5×5 subset of the whole heat map are shown for better visualisation. The complete matrices of estimates and the truth are quite similar.

Table 1 summarises simulation results when the data generating process is BMC. Only the results from BMC$_j$ are presented because it showed slightly improved performance over the other two. Note that Area Under the ROC Curves (AUCs) from tcpl in Table 1 & S1 were computed slightly differently than those from other methods. BMC, three variations, and ZIPLL all produce *probability* of active responses, which can be any value between 0 and 1. In order to evaluate the accuracy of estimates compared to the true $\gamma_{ij} \in \{0, 1\}$ values, ROC curves and the corresponding AUCs are computed by changing thresholds between 0 and 1. On the other hand, EPA provides a *binary* hit-call variable

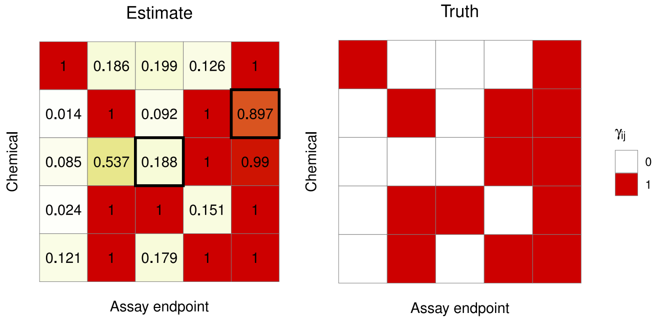

Fig. 3. Heat map of estimated and true profiles of the mean effect from Simulation 1. Figure presents the results from a 5×5 subset chosen at random. The value in each cell of the left panel is the posterior mean of γ_{ij}. Cells with outer lines ((3,3) and (2,5) elements) are hold-out pairs for which γ_{ij}'s are predicted.

for the mean effect through ToxCast/Tox21. We hereafter refer to this variable (based on the version invitroDBv2) as EPA's hit-call. The EPA's hit-call identifies a pair as active if the fitted Hill or gain-loss model have lower Akaike information criterion than a constant model, and both the estimated and observed maximum responses exceed a cutoff chosen for the assay endpoint. This classification of whether each pair is active or not is directly comparable to the true γ_{ij} without changing thresholds. In simulations, assay endpoint-specific cutoffs are set to one standard deviation away from the median of responses for that assay.

Table 1 shows that BMC outperforms the other methods overall. In training data sets, BMC approaches (BMC and BMC$_j$) have lower RMSEs and higher AUCs compared to tcpl or ZIPLL. Poor performance of ZIPLL in these simulations is partially due to the facts that monotone increasing shape restrictions fail to fit decreasing trends and that ZIPLL does not allow for different σ_j^2's. BMC outperforming tcpl may be due to the borrowing of information across chemicals and assay endpoints. Moreover, the original BMC model produced in- and out-of-sample AUCs that are uniformly better than those from BMC$_j$. Hence, when the factor model provides a realistic characterisation of the dependence structure across assay endpoints and chemicals, it is not suggested to use a simplified model for multiple testing. Less structure in γ_{ij} results in lower out-of-sample

Table 1. Summary of results from Simulation 1. Root Mean Squared Error (RMSE), Area Under the ROC curve (AUC) results for probability of the mean effect, and AUC for probability of the variance effect are presented. The displayed values are the mean (standard error) across 50 simulations.

	BMC	BMC$_j$	ZIPLL	tcpl
RMSE	0.402 (0.018)	0.397 (0.017)	0.820 (0.034)	0.645 (0.017)
In-sample AUC for γ_{ij}	0.997 (0.001)	0.996 (0.001)	0.661 (0.008)	0.856 (0.007)
Out-of-sample AUC for γ_{ij}	0.762 (0.085)	0.506 (0.047)	-	-
In-sample AUC for t_{ij}	1.000 (0.000)	1.000 (0.000)	-	-

AUCs for γ_{ij} (BMC$_0$: 0.504 (0.054), BMC$_i$: 0.486 (0.049)), where BMC$_0$ does not utilise chemical- or assay endpoint-specific information in prediction.

Another benefit of BMC is the capability of modelling heteroscedasticity. The AUCs for t_{ij} in Table 1 exhibit highly accurate estimation performance for the probability of heteroscedastic variances. Figure 2 illustrates that BMC closely recovers the true curves even in the existence of heteroscedasticity (A and B) and that it nicely differentiates variance changes and mean changes - for instance, the estimated probability of the variance effect is around 0.1, while the probability of the mean effect is 1 in C. Figure S7 in the Supporting Material shows residuals versus fitted values for the heteroscedastic pairs in Figure 2. The ZIPLL does not consider heteroscedasticity in the model and consequently results in heteroscedastic residuals. In contrast, BMC is able to properly account for heteroscedasticity, and residuals do not show any patterns against fitted values.

Simulation 2 generates data from an alternative model, ZIPLL. Despite misalignment in data structure assumed by BMC and by ZIPLL, BMC performs similarly to ZIPLL and outperforms tcpl with respect to RMSE and AUC. The high in-sample AUC for γ_{ij} from BMC (0.982) suggests its stable estimation performance even with relatively small number of chemicals and assay endpoints. We provide a full discussion of Simulation 2 results in the Supporting Material S3.

5.2. ToxCast/Tox21 Results

This section presents results from the ToxCast/Tox21 data analysis with a focus on endpoints relevant to human neurodevelopmental disorders and obesity. We ran the sampler for 40,000 iterations from which 30,000 were discarded as burn-in, and every 10th sample was saved for the next 10,000 iterations. This long burn-in is to be conservative; trace plots and effective sample sizes for MCMC samples indicated good mixing and apparent convergence after 15,000 iterations.

Figures 4 show estimated dose-response curves from BMC as dashed lines with 95% credible intervals as shaded areas with darker gray. The lighter gray shaded areas illustrate 95% predictive intervals for the data points drawn as dots. “Pr(Mean Effect)” is the mean effect probability for a (chemical i, assay endpoint j) pair, which is computed as the posterior mean of γ_{ij}. Similarly, “Pr(Var Effect)” means the variance effect probability whose value is the posterior mean of t_{ij}. The posterior mean of π_t is 0.12, meaning that about 12% of the observed cells have heteroscedastic variances.

The first row of Figure 4 shows that BMC is able to differentiate dose effects on the mean from those on the variance of dose-response curves. Recall that the EPA’s hit-call is an indication of mean changes. In the left panel, BMC and the EPA agree that mean changes exist, which is supported by an increasing trend. In the right panel, the EPA’s hit-call claims that the average dose-response is not constant. However, BMC estimates that the mean curve is constant at zero with probability 0.75, but with there being clear evidence of heteroscedasticity. Therefore, the first row in Figure 4 suggests that (1) the EPA’s hit-call for the mean effect might be misled by heteroscedastic variances; and (2) BMC can separate mean and variance effects (at least in some cases).

The second row of Figure 4 illustrates some cases where BMC and the EPA’s hit-call disagree, and BMC’s result is more plausible. The EPA’s hit-calls say no activity in the mean for both plots, while BMC estimates them to be active with high probability. In these cases, not only do plots show increasing trends of dose-response measurements, but also background knowledge supports BMC’s estimates. In fact, Bisphenol A (BPA) and phthalates are known to disrupt the endocrine system, which potentially results in neurodevelopmental disorders (Tran and Miyake, 2017) and obesity (Holtcamp,

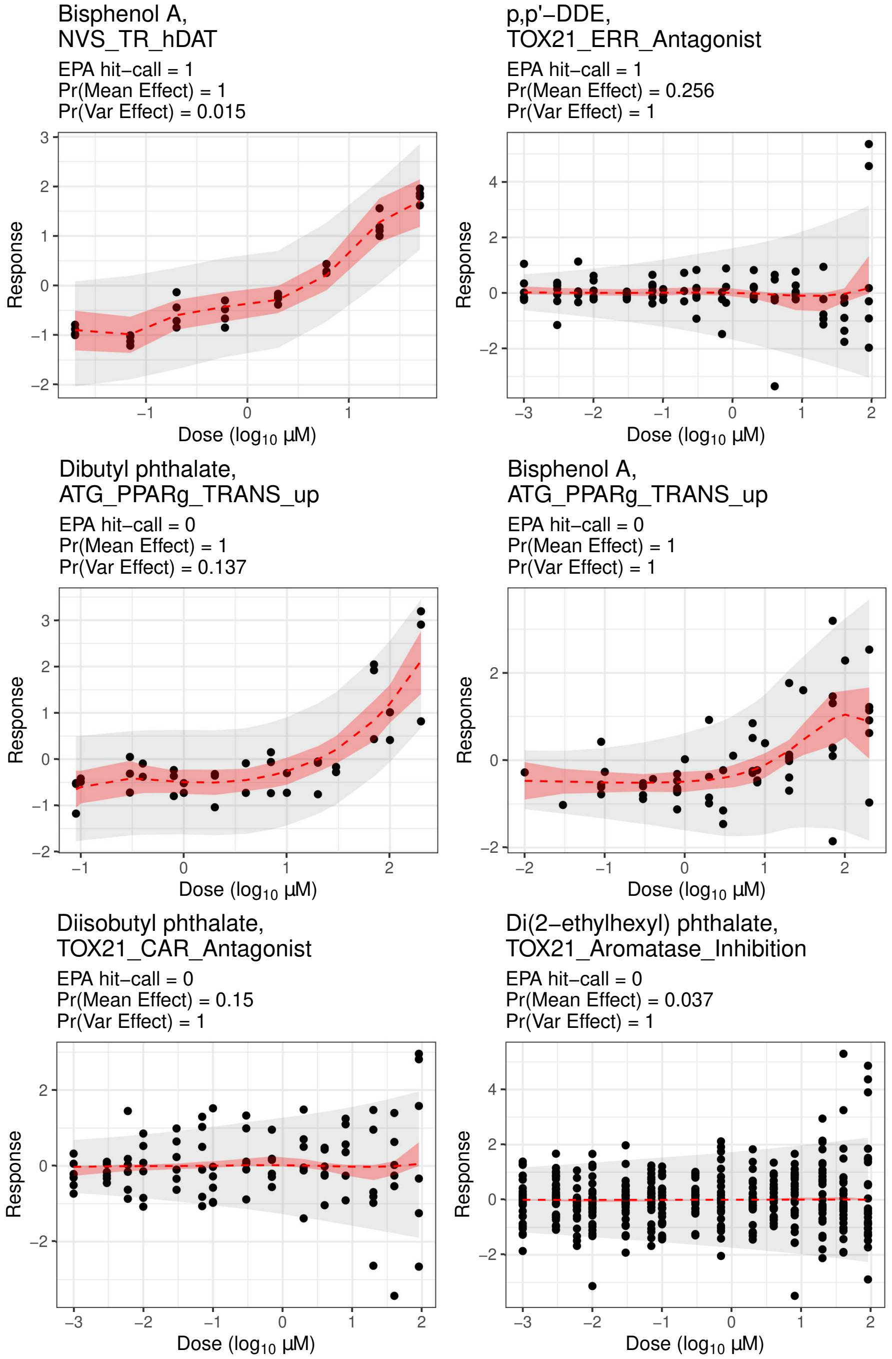

Fig. 4. Fitted results for select chemical-assay endpoint pairs estimated to be active by BMC. The first row shows pairs having dose effects on the mean (left) or variance (right). The second row illustrates dose effects on the mean in both panels. The third row presents pairs both having dose effects on the variance.

2012). In addition, chemicals activating PPAR$_\gamma$ receptors are potential obesogens because PPAR$_\gamma$ is a master regulator in formulating fat cells (Evans et al., 2004). Therefore, it is not unexpected for BPA and Dibutyl phthalate to be active for the assay endpoint, ATG_PPAR$_\gamma$_TRANS_up. Several pieces of evidence reinforce the validity of results from BMC over the EPA's hit-call.

The third row of Figure 4 shows cases where the EPA's hit-call can have low power because it misses signals manifest in the variance instead of the mean. Given that phthalates are related to obesity (Holtcamp, 2012), we expect disruptive patterns on assay endpoints presenting toxicity of Diisobutyl phthatlate and Di(2-ethylhexyl) phthatlate. However, the EPA's hit-call suggests that these phthalates are not active at the doses tested. This may be true in terms of mean changes, but variances seem clearly heteroscedastic.

Figures 5 and S8 show predicted results for hold-out pairs. Note that we do not predict observations or the average dose-response curves. BMC only predicts the mean effect probability, which is often the primary focus of many studies and helps researchers to prioritise chemicals for further testing.

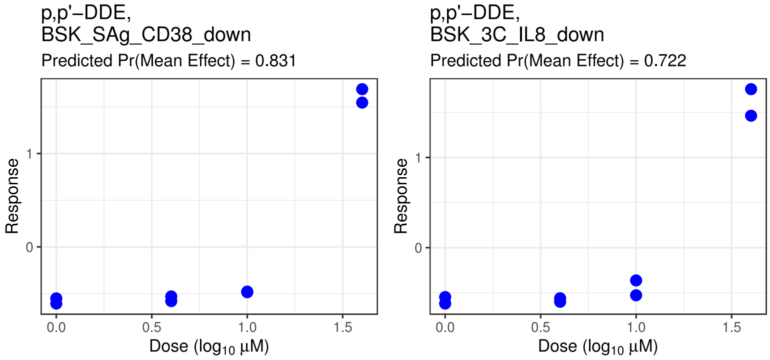

Fig. 5. Results for select chemical-assay endpoint pairs with similar measurements but predicted by BMC to have different probabilities of the mean effect.

Figure 5 displays advantages of BMC's probabilistic approach to evaluate dose effects on the mean. The left and right panels exhibit almost identical dose-response results of the same chemical and different assay endpoints. These pairs have different mean effect probabilities that reflect different assay endpoint effects. The chemical p,p'-DDE is predicted to cause mean changes in dose-response curves with average probability of 0.808 across assay endpoints. Across chemicals, the assay endpoint BSK_SAg_CD38_down (left) is more likely to have the mean effect with the average posterior probability 0.874 than BSK_3C_IL8_down (right) with 0.694. Hence, the predicted probabilities for the mean effect can be thought to be pulled towards the probability of each assay endpoint from the chemical's probability. This implies that BMC appropriately addresses chemical and assay endpoint effects in the mean effect probabilities through the latent factor model. On the other hand, the EPA's hit-call is 1 for both cases. Their deterministic approach might not always be informative when researchers attempt to arrange chemicals by evidence of toxicity. When the researchers are more informed by the probabilities, however, they can easily prioritise chemicals - even choosing among those with the same hit-call.

We observed that seemingly active pairs with mean changes can have a wide range of mean effect probabilities. (Refer to Figure S8 in the Supporting Material for some relatively active pairs.) In fact, we found that in the ToxCast/Tox21 application, 95% highest density intervals for the estimated and predicted γ_{ij}'s are (0.205, 1) and (0.510, 0.863), respectively. These suggest the lack of conclusive evidence of inactivity in most cases, while the EPA's hit-call forces chemical-assay endpoint pairs to be classified as active or inactive. The EPA's hit-call may tend to flag too many pairs as inactive.

It is valuable to assess toxicity of chemicals based on their activity probability, which could be computed as $Pr(\gamma_{ij} = 1 \cup t_{ij} = 1) = 1 - Pr(\gamma_{ij} = 0 \cap t_{ij} = 0)$. Figure S9 in the Supporting Material shows chemicals by the order of average activity probability over obesity-related assay endpoints. Top chemicals that are most likely to disrupt biological processes associated with obesity include p,p'-DDE, Dichlorodiphenyltrichloroethane (DDT), Triclosan, BPA, 2,4,5-Trichlorophenol, Chlorpyrifos, and Benzyl butyl phthalate. The rankings of chemicals by BMC and by the EPA's hit-call show subtle differences.

Chemicals are similarly ranked by the average probability over assay endpoints related to neurodevelopmental disorders (Figure S10 in the Supporting Material). Top chemicals that are most likely to disrupt neurodevelopmental processes include Triclosan, DDT, 2,4,5-Trichlorophenol, p,p'-DDE, Fenpropathrin, 2-Hydroxy-4-methoxybenzophenone, and BPA. Between the two sets of the most active chemicals associated with neurodevelopmental disorders and obesity, five chemicals - Triclosan, BPA, 2,4,5-Trichlorophenol, DDT, and p,p'-DDE - overlap, and we call them the top 5 chemicals. These bioactivity rankings are based on the data that are currently available. As data expand, it will be informative to revisit such rankings.

Figure S11 in the Supporting Material provides a list of assay endpoints of relevance to neurodevelopmental disorders, which are highly likely (> 0.9) to be "activated" by the top 5 chemicals. The list includes both agonist and antagonist assays, and thus the "activated" probability encompasses agonist and antagonist directions in the mean effect. The assay endpoints in the list are expected to have important implications in disease progression, from which thirty-one assay endpoints show impacts on both disease classes. The same list for the obesity-related assay endpoints is also provided in the Supporting Material (Figure S12).

To study sensitivity of rankings to the choice of chemicals, we expanded our analysis to 326 chemicals. They consist of the original 30 chemicals and those screened in Phase I of the ToxCast that have been exclusively used in other toxicity studies including Martin et al. (2010) and Wilson et al. (2014). Within this larger collection, relative positions of the 30 chemicals remained intact with a few exceptions. BPA and Triclosan were positioned lower in the larger set, while Cyfluthrin and MEHP were positioned higher. One of explanations for these shifts is an altered correlation structure among chemicals. The Phase I chemicals are mostly pesticides, and the four chemicals might have different relationship with those from what they had with the 30 chemicals in terms of the mean effect.

6. Discussion

We have proposed a Bayesian multiple testing approach for inference on activity of chemicals in settings involving multiple chemicals and assay endpoints and possible heteroscedasticity. Our BMC approach can be applied directly in other settings involving a similar matrix-structured experimental design. For example, this is common in pharmaceutical studies assessing drug activity - studies will look for evidence of activity for different health outcomes. Also, in microbial genetics, similar designs are conducted but for different types of bacteria and environmental conditions.

The ultimate goal of many analyses using *in vitro* data is to make inferences on human health and inform protective regulations. Accordingly, chemicals and assay endpoints studied in the ToxCast/Tox21 application are carefully selected: the chemicals are also measured in human epidemiology studies, and the assay endpoints cover a variety of species and several types of tissue targets. It will be interesting to follow up on the top ranking chemicals for neurodevelopmental disorders and obesity outcomes identified in our analyses to further elucidate their role in human health.

When extending *in vitro* results to *in vivo* toxicity, doses need to be carefully considered. All the results presented in the paper should be interpreted in terms of tested doses, so we do not conclude a chemical with a high probability of inactivity is inactive at higher doses than those tested. Simultaneously, it is recommended to ensure that the doses tested *in vitro* can physiologically occur in animals/humans. This recommendation is reinforced by Klaren et al. (2019) in which *in vivo* toxicity prediction using *in vitro* assays performs much better with toxicokinetic modelling. Therefore, future research linking *in vitro* data and *in vivo* implications could be greatly assisted by assuring dose applicability in animals/humans as well as widening the range of tested doses.

Acknowledgments

We are grateful for the financial support of the National Institute of Environmental Health Sciences through grants R01ES027498 and R01ES028804. The authors would like to thank Brett Winters for identifying assay endpoints relevant to neurodevelopmental disorders and obesity, Evan Poworoznek for sharing computer code, Kelly Moran for helpful comments, and Matthew Wheeler for help in processing ToxCast/Tox21 data.

References

Bhattacharya, A. and Dunson, D. B. (2011) Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

Corty, R. W. and Valdar, W. (2018) Vqtl: an R package for mean-variance QTL mapping. *G3: Genes, Genomes, Genetics*, **8**, 3757–3766.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T. C. and Mattingly, C. J. (2019) The comparative toxicogenomics database: update 2019. *Nucleic Acids Research*, **47**, D948–D954.

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W. and Kavlock, R. J. (2007) The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences*, **95**, 5–12.

Durante, D. (2017) A note on the multiplicative gamma process. *Statistics & Probability Letters*, **122**, 198–204.

ECHA (2017) The use of alternatives to testing on animals for the REACH regulation. *European Chemicals Agency, Helsinki, Finland*, `https://doi.org/10.2823/023078`.

Evans, R. M., Barish, G. D. and Wang, Y.-X. (2004) PPARs and the complex journey to obesity. *Nature Medicine*, **10**, 355–361.

Filer, D. L., Kothiya, P., Setzer, R. W., Judson, R. S. and Martin, M. T. (2017) Tcpl: the ToxCast pipeline for high-throughput screening data. *Bioinformatics*, **33**, 618–620.

Holtcamp, W. (2012) Obesogens: an environmental link to obesity. *Environmental Health Perspectives*, **120**, a62–a68.

Judson, R., Houck, K., Martin, M., Richard, A. M., Knudsen, T. B., Shah, I., Little, S., Wambaugh, J., Woodrow Setzer, R., Kothya, P. et al. (2016) Editor's highlight: analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space. *Toxicological Sciences*, **152**, 323–339.

Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., Reif, D. M., Rotroff, D. M., Shah, I., Richard, A. M. et al. (2010) In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environmental Health Perspectives*, **118**, 485–492.

Klaren, W. D., Ring, C., Harris, M. A., Thompson, C. M., Borghoff, S., Sipes, N. S., Hsieh, J.-H., Auerbach, S. S. and Rager, J. E. (2019) Identifying attributes that influence in vitro-to-in vivo concordance by comparing in vitro Tox21 bioactivity versus in vivo drugmatrix transcriptomic responses across 130 chemicals. *Toxicological Sciences*, **167**, 157–171.

Knapen, D., Stinckens, E., Cavallin, J. E., Ankley, G. T., Holbech, H., Villeneuve, D. L. and Vergauwen, L. (2020) Toward an AOP network-based tiered testing strategy for the assessment of thyroid hormone disruption. *Environmental Science & Technology*, **54**, 8491–8499.

Koren, Y., Bell, R. and Volinsky, C. (2009) Matrix factorization techniques for recommender systems. *Computer*, **42**, 30–37.

Leslie, D. S., Kohn, R. and Nott, D. J. (2007) A general approach to heteroscedastic linear regression. *Statistics and Computing*, **17**, 131–146.

Li, F. and Zhang, N. R. (2010) Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, **105**, 1202–1214.

Martin, M. T., Dix, D. J., Judson, R. S., Kavlock, R. J., Reif, D. M., Richard, A. M., Rotroff, D. M., Romanov, S., Medvedev, A., Poltoratskaya, N. et al. (2010) Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's toxcast program. *Chemical Research in Toxicology*, **23**, 578–590.

Mnih, A. and Salakhutdinov, R. R. (2008) Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, 1257–1264.

Neelon, B. and Dunson, D. B. (2004) Bayesian isotonic regression and trend analysis. *Biometrics*, **60**, 398–406.

Paré, G., Cook, N. R., Ridker, P. M. and Chasman, D. I. (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the women's genome health study. *PLoS Genetics*, **6**.

Purushotham, S., Liu, Y. and Kuo, C.-C. J. (2012) Collaborative topic regression with social matrix factorization for recommendation systems. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 691–698.

Ritz, C. (2010) Toward a unified approach to dose–response modeling in ecotoxicology. *Environmental Toxicology and Chemistry*, **29**, 220–229.

Rönnegård, L. and Valdar, W. (2012) Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genetics*, **13**, 63.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M. and Meze-Hausken, E. (2013) A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**, 85–100.

Scott, J. G. and Berger, J. O. (2006) An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, **136**, 2144–2162.

— (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, **38**, 2587–2619.

Tansey, W., Tosh, C. and Blei, D. M. (2019) Relational dose-response modeling for cancer drug studies. *arXiv preprint*, `https://arxiv.org/abs/1906.04072v2`.

Thomas, D. C., Conti, D. V., Baurley, J., Nijhout, F., Reed, M. and Ulrich, C. M. (2009) Use of pathway information in molecular epidemiology. *Human Genomics*, **4**, 21.

Tice, R. R., Austin, C. P., Kavlock, R. J. and Bucher, J. R. (2013) Improving the human hazard characterization of chemicals: a tox21 update. *Environmental Health Perspectives*, **121**, 756–765.

Tran, N. Q. V. and Miyake, K. (2017) Neurodevelopmental disorders and environmental toxicants: Epigenetics as an underlying mechanism. *International Journal of Genomics*, **2017**, 1–23.

Wilson, A., Reif, D. M. and Reich, B. J. (2014) Hierarchical dose–response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics*, **70**, 237–246.

Yang, J., Loos, R. J., Powell, J. E., Medland, S. E., Speliotes, E. K., Chasman, D. I., Rose, L. M., Thorleifsson, G., Steinthorsdottir, V., Mägi, R. et al. (2012) FTO genotype is associated with phenotypic variability of body mass index. *Nature*, **490**, 267–272.

Web-based supporting materials for “Bayesian Matrix Completion for Hypothesis Testing”

S1. Figures

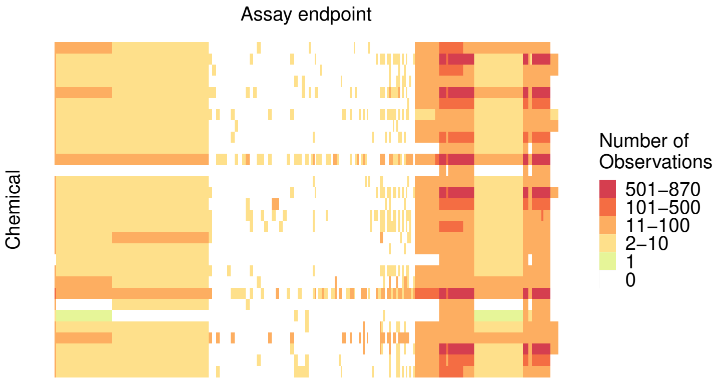

Fig. S1. Heat map of the number of observations in ToxCast/Tox21 data for obesity, based on 30 chemicals (rows) and 271 assay endpoints (columns).

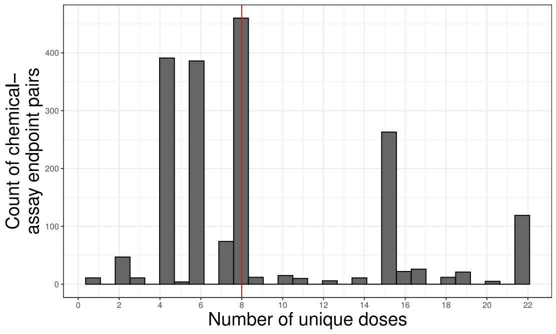

Fig. S2. Histogram of the number of unique doses tested for each chemical-assay endpoint pair in the ToxCast/Tox21 data related to neurodevelopmental disorders. The vertical solid line is the median. Half the pairs have the number of tested doses less than or equal to 8.

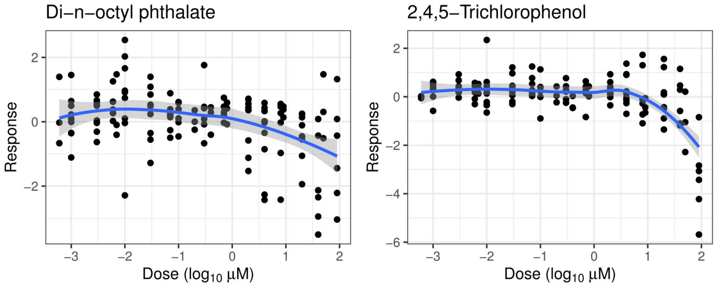

Fig. S3. Scatter plots of two chemicals on the TOX21_ERa_LUC_BG1_Agonist assay endpoint. The solid lines and gray shaded areas represent the average dose-response curves and 95% confidence intervals fitted via locally estimated scatterplot smoothing (LOESS).

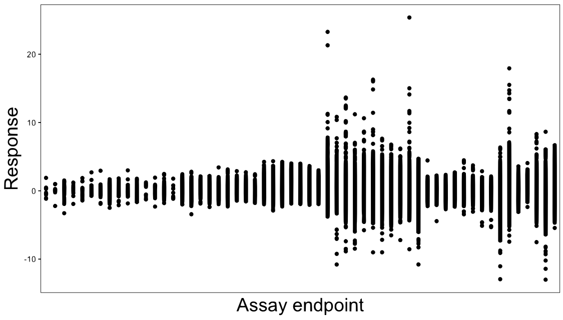

Fig. S4. Scatter plot of the responses normalised by chemical-assay endpoint pairs. Points on each vertical line represent responses from one assay endpoint. This figure is based on a subset of assay endpoints.

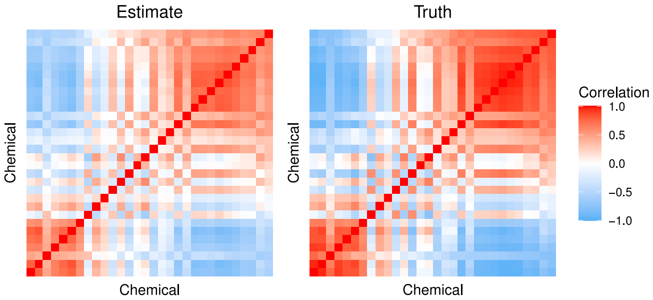

Fig. S5. Heat map of the estimated and true correlation matrix among chemicals with respect to the mean effect. The results are from Simulation 1.

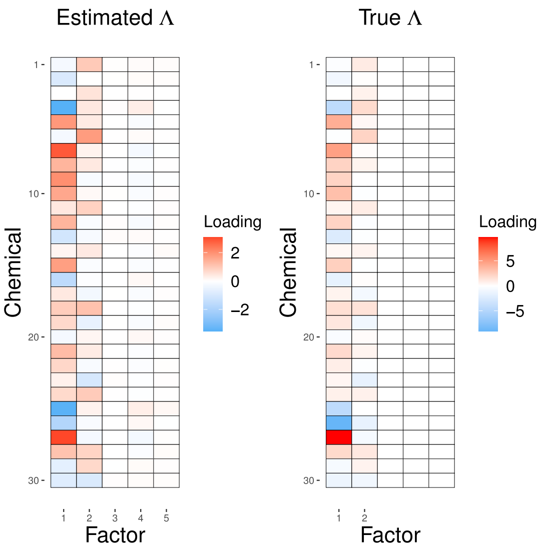

Fig. S6. The estimated and true entries of loading matrix Λ from Simulation 1. The estimated Λ is rotated for better visualization.

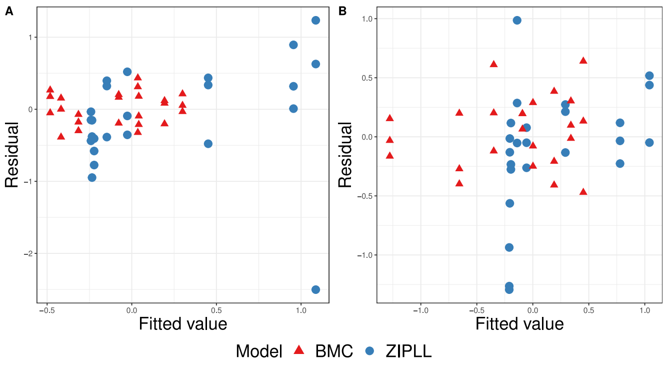

Fig. S7. Residuals versus fitted values using BMC and ZIPLL in Simulation 1. The residuals and fitted values in A and B are computed using observations and fitted lines from A and B in Figure 2, respectively. Note that residuals from ZIPLL are obtained by subtracting the fitted values from observations, while those from BMC are the posterior mean of “normalised” residuals whose value at sth iteration is $(y_{ijk} - \gamma_{ij}^{(s)} f_{ij}^{(s)}(x_{ijk}))/\exp(x_{ijk}\delta_{ij}^{(s)}/2)$.

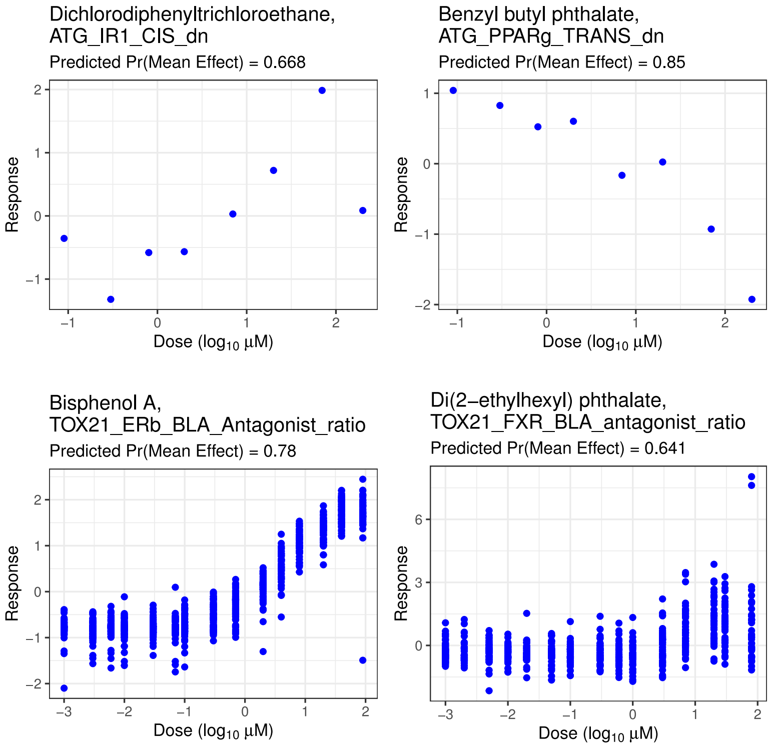

Fig. S8. Results for select chemical-assay endpoint pairs predicted by BMC to likely have dose effects on the mean.

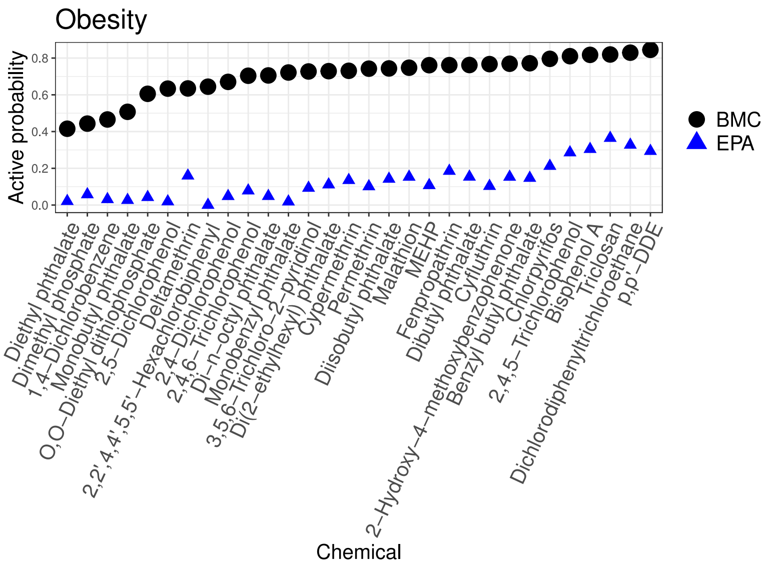

Fig. S9. Chemical ranks by the average activity probability from BMC (dots) and the average hit-call from EPA (triangles) over obesity-related assay endpoints.

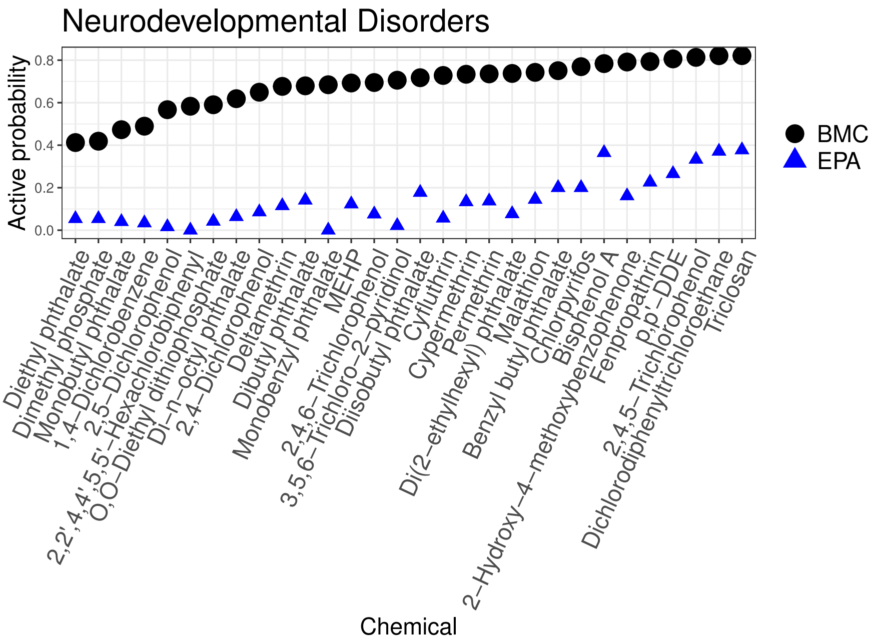

Fig. S10. Chemical ranks by the average active probability from BMC (dots) and the average hit-call from EPA (triangles) over assay endpoints related to neurodevelopmental disorders.

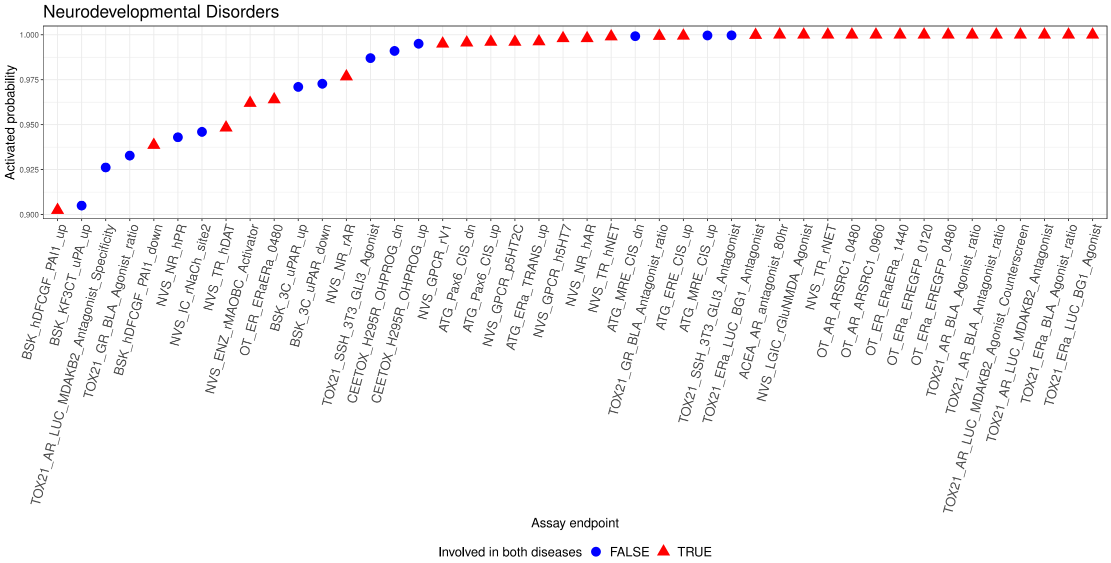

Fig. S11. Ranks of assay endpoints associated with neurodevelopmental disorders in terms of probabilities to be activated by the top 5 chemicals. Only a subset of assay endpoints have activation probabilities higher than 0.9. The assay endpoints with dots are marked uniquely for neurodevelopmental disorders, while those with triangles are marked for two disease classes.

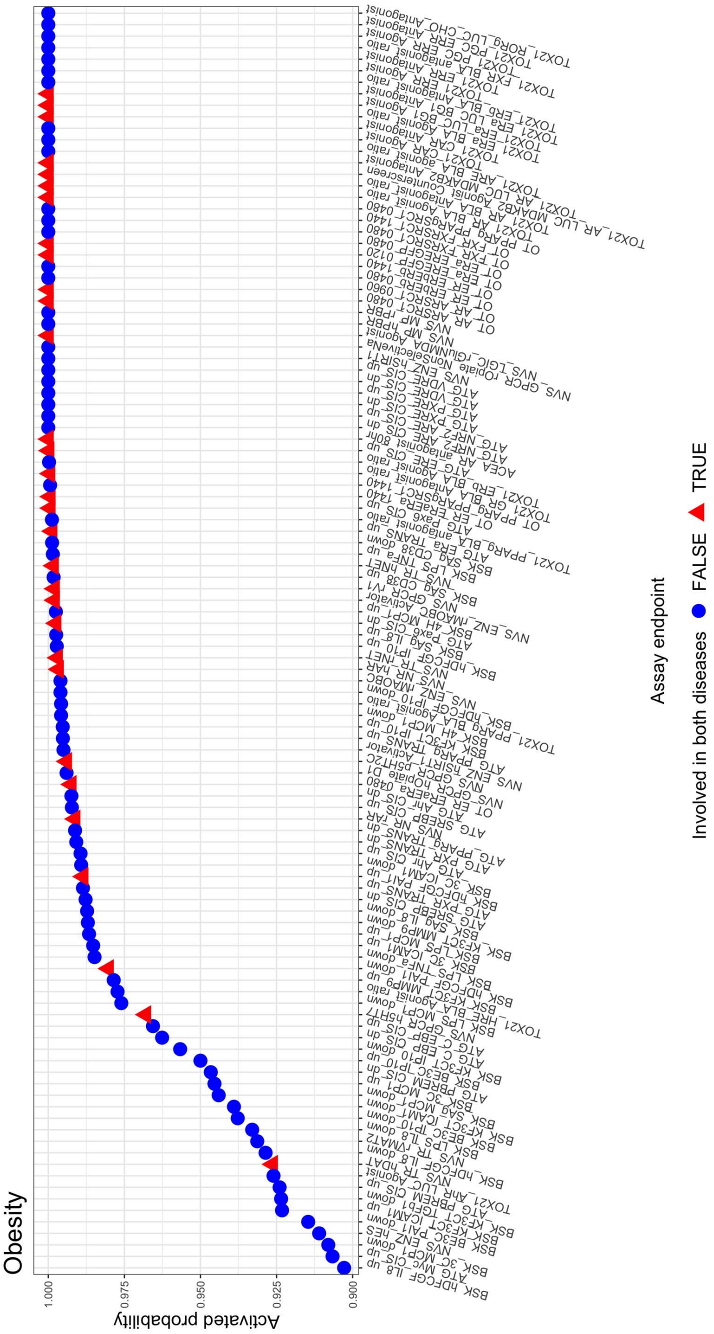

Fig. S12. Ranks of obesity-related assay endpoints in terms of probabilities to be activated by the top 5 chemicals. Only a subset of assay endpoints are presented with the activated probabilities higher than 0.9. The assay endpoints with dots are marked uniquely for obesity, while those with triangles are marked for both diseases.

S2. Data

This section explains our selection criteria for chemicals and assay endpoints related to neurodevelopmental disorders and obesity in the ToxCast/Tox21 data. The exact procedure to find assay endpoints of interest is as follows: First, molecules are identified that have known associations with each disease through the Comparative Toxicogenomics Database (CTD) (Davis et al., 2019) and Ingenuity® PathwayAnalysis (IPA) Knowledgebase (QIAGEN Inc., `https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/`). Here, the known associations include: molecules are biomarkers of the disease; are known to play a role in the disease etiology; and are therapeutic targets for treatment of the disease. The databases CTD and IPA maintain curated and published associations between molecules and diseases. It is noteworthy that they originated from a variety of species and tissue targets. In later steps, the databases are compared to the ToxCast/Tox21 data, whose assay endpoints were derived from different species. Moreover, the assay endpoints in the ToxCast/Tox21 program were tested across several types of tissue targets that may differently mediate the relationship of even the same molecular target and the same assay endpoint. Therefore, it is important to ensure that a wide variety of species and tissue targets are well represented in both ToxCast/Tox21 and the databases from which molecular targets are identified. Second, we filter molecular targets in ToxCast/Tox21 that overlap with the identified molecules from CTD and IPA. Third, we choose assay endpoints that those overlapping molecular targets are screened over. As a result of these steps, 132 and 352 assay endpoints were identified as relevant to neurodevelopmental disorders and obesity, respectively, which were further filtered based on chemical coverage, as detailed below.

A partial list of chemicals is considered due to a particular interest in human data. We featured a set of overlapping chemicals measured in ToxCast/Tox21 and an existing observational study of environmental risk factors for neurodevelopmental disorders and obesity. In doing so, we believe future application of the ToxCast/Tox21 results to humans will be more viable. A total of 48 chemicals were selected, 30 of which were tested within the above mentioned list of assay endpoints. Due to the reduced list of chemicals, the number of assay endpoints has diminished as well. Following recommended practice

(Judson et al., 2016), we retained only the doses lower than a cytotoxicity point for each chemical, which removed two percent of the data. We employed the cytotoxicity median values stored in a variable "cyto_pt_um" in the tcpl package (Filer et al., 2017). Consequently, our final data involve 30 chemicals and 131 and 271 assay endpoints related to neurodevelopmental disorders and obesity, respectively.

S3. Simulations

For all simulations, we used eight unique doses {0.301, 0.477, 0.602, 0.845, 1.000, 1.301, 1.602, 2.000} in $\log_{10}$ μM chosen based on the frequency of appearance in the ToxCast/Tox21 data. B-spline knots are set at the minimum value, three quartiles, and the maximum of the doses.

In Simulation 1, we generated 50 data sets with $K_{ij} = 24$ at each combination of chemical and assay endpoint, which represents 3 replicates at each of the eight doses. Elements of $\boldsymbol{\eta}$ were drawn independently from the standard normal distribution. Elements in the $m \times q$ matrix Λ were sampled as in (3) and (4) with $\nu = 3$, $a_1 = 2.1$, and $a_2 = 3.1$, following the note by (Durante, 2017). The assay endpoint-specific variances were sampled from $1/\sigma_j^2 \sim Gamma(\frac{5}{2}, \frac{5\times0.1}{2})$. Heteroscedasticity is expected at one fifth of chemical-assay endpoint combinations by setting $\pi_t = 0.2$. For heteroscedastic pairs, d_{ij} was sampled from $N(2, 0.1^2)$, which gives roughly $\exp(x_{ijk})\epsilon_{ijk}$ for the error term. For BMC and ZIPLL, 20,000 samples were drawn, of which 1,000 samples were saved and analysed. First 10,000 samples were discarded as burn-in, and every 10th sample was retained for the next 10,000 samples. Trace plots and effective sample sizes for MCMC samples suggested convergence and good mixing.

In Simulation 2 where misalignment exists between the true data generating process and BMC, we generated data under the ZIPLL model. The number of chemicals m was set to 15, and the number of assay endpoints J to 15. The bottom right 3×3 chemical-assay endpoint pairs were held out for prediction of γ_{ij}. We generated 50 data sets with $K_{ij} = 8$ so that each chemical-assay endpoint pair has one observation at each of the eight doses. No replicates at such scarce doses make it impractical to evaluate

Table S1. Summary of results from Simulation 2. The RMSEs and AUC results of the mean effect probabilities are presented. The displayed values are the mean (standard error) across 50 simulated data sets.

	BMC	BMC$_0$	ZIPLL	tcpl
RMSE	0.094 (0.002)	0.094 (0.002)	0.088 (0.002)	0.282 (0.017)
In-sample AUC for γ_{ij}	0.982 (0.009)	0.983 (0.008)	0.981 (0.008)	0.907 (0.018)
Out-of-sample AUC for γ_{ij}	0.506 (0.228)	0.475 (0.209)	-	-

heteroscedasticity, which is consequently not considered in Simulation 2. The model

$$y_{ijk} = \gamma_{ij} f_{ij}(x_{ijk}) + \epsilon_{ijk},\ \epsilon_{ijk} \sim N(0, 0.1^2)$$

was considered where around half the pairs were randomly assigned to have the mean effect with $\gamma_{ij} \sim Bernoulli(0.5)$. The dose-response function was

$$f_{ij}(x_{ijk}) = t_{ij} - \frac{t_{ij} - b_{ij}}{1 + \exp\{w_{ij}(\log x_{ijk} - \log a_{ij})\}}$$

where $t_{ij} \sim Unif(0, 10)$, $b_{ij} = 0$, $a_{ij} = max(x_{ijk})$, and $w_{ij} \sim Unif(1, 8)$. The RMSE and AUC results are summarised in Table S1. Only the results from BMC$_0$ are presented because ZIPLL adopts BMC$_0$ in its γ_{ij} testing framework.

It is remarkable that BMC is able to estimate dose-response trends almost as well as ZIPLL and has similar accuracy in estimating γ_{ij} even when ZIPLL is the true data generating process. In Simulation 2, BMC and ZIPLL outperform tcpl models. Smaller RMSEs and higher AUCs from BMC and ZIPLL compared to those from tcpl suggest increased robustness of spline methods than parametric ones for dose-response functions. The improved metrics also indicate benefits of hierarchical methods over the independent curve fitting that ignores correlations between chemicals or assay endpoints. In particular, the achievement of the high in-sample AUC for γ_{ij} from BMC is encouraging despite the relatively small number of chemicals and assay endpoints. The disadvantage of these small m and J, however, is apparent in poor predictive AUCs of BMC and BMC$_0$. The out-of-sample AUC for γ_{ij} from BMC is only slightly better than random guessing (AUC = 0.5), while the models with simpler structures on γ_{ij} are not as good as random guessing (BMC$_i$: 0.475 (0.199), BMC$_j$: 0.5 (0.242)).

S4. Posterior Computation

Under the prior specification in section 4, posterior samples are obtained by iterating the following partially collapsed MCMC sampler.

Heteroscedasticity

(a) Update t_{ij} and δ_{ij} simultaneously using the Metropolis-Hastings algorithm. Propose t^p_{ij} as follows: for each j, choose random number of elements and random indices to update. For those selected (i,j) pairs, flip zero and one. Given the proposed t^p_{ij}, propose δ^p_{ij} using t-distribution with 4 degrees of freedom centered at the current δ^c_{ij}. Accept $(t^p_{ij}, \delta^p_{ij})$ with probability $min\{1, r\}$ where

$$r = \frac{\prod_{k=1}^{K_{ij}} N(y_{ijk} - \gamma_{ij}(\mathbf{x}^B_{ijk})^T \boldsymbol{\beta}_{ij}; 0, \exp(x_{ijk}\delta^p_{ij}/2)^2\sigma_j^2)}{\prod_{k=1}^{K_{ij}} N(y_{ijk} - \gamma_{ij}(\mathbf{x}^B_{ijk})^T \boldsymbol{\beta}_{ij}; 0, \exp(x_{ijk}\delta^c_{ij}/2)^2\sigma_j^2)} \times \\ \frac{Bernoulli(t^p_{ij}; \pi_t)\{N(\delta^p_{ij}|t^p_{ij} = 1; 0, v_\delta)\mathbf{1}(t^p_{ij} = 1) + 1 \times \mathbf{1}(t^p_{ij} = 0)\}}{Bernoulli(t^c_{ij}; \pi_t)\{N(\delta^c_{ij}|t^c_{ij} = 1; 0, v_\delta)\mathbf{1}(t^c_{ij} = 1) + 1 \times \mathbf{1}(t^c_{ij} = 0)\}}$$

and $(\mathbf{x}^B_{ijk})^T$ is the kth row of the B-spline basis matrix X_{ij}.

(b) Update π_t from

$$(\pi_t | t_{ij}\ \forall i, j) \sim Beta\left(1 + n_t, 1 + \sum_{j=1}^{J} m_j - n_t\right),$$

where $n_t = \sum_i \sum_j \mathbf{1}(t_{ij} = 1)$ and $m_j = \sum_i \mathbf{1}(K_{ij} > 0)$.

Functional Mean

Once Steps 1&2 are completed in every iteration, the data (X, Y) need to be reformulated: y_{ijk} is replaced by $y_{ijk}/\exp(x_{ijk}\delta_{ij}/2)$, and $\mathbf{x}^B_{ijk}$ by $\mathbf{x}^B_{ijk}/\exp(x_{ijk}\delta_{ij}/2)$.

(c) Update $(\boldsymbol{\lambda}_i, \boldsymbol{\eta}_j)$ as in (Bhattacharya and Dunson, 2011). Priors, hyperparameters specification and posterior distributions are fully explained in (Bhattacharya and Dunson, 2011) and (Durante, 2017). We shall not repeat the sampling algorithms for $\boldsymbol{\lambda}_i$ and $\boldsymbol{\eta}_j$ here. With the sampled $(\boldsymbol{\lambda}_i, \boldsymbol{\eta}_j)$, update π_{ij} using (1).

(d) Update z_{ij} from

$$(z_{ij}|\gamma_{ij}=1,\boldsymbol{\lambda}_i,\boldsymbol{\eta}_j)\sim TN_{(0,\infty)}(\boldsymbol{\lambda}_i^T\boldsymbol{\eta}_j,1),$$
$$(z_{ij}|\gamma_{ij}=0,\boldsymbol{\lambda}_i,\boldsymbol{\eta}_j)\sim TN_{(-\infty,0)}(\boldsymbol{\lambda}_i^T\boldsymbol{\eta}_j,1)$$

where $TN_{(a,b)}(\mu,\sigma^2)$ denotes a normal distribution truncated to the interval (a,b) with mean μ, variance σ^2.

(e) Update γ_{ij} from the conditional Bernoulli distribution with $\boldsymbol{\beta}_{ij}$ marginalised out. Using the following probabilities,

$$Pr(\gamma_{ij}=1|\mathbf{y}_{ij},X_{ij},\sigma_j^2,\Sigma_j,\pi_{ij})$$
$$\propto \pi_{ij}|\Sigma_j X_{ij}^T X_{ij}/\sigma_j^2+I_p|^{-1/2}\times\exp\left(\frac{1}{2\sigma_j^4}\mathbf{y}_{ij}^T X_{ij}\left(X_{ij}^T X_{ij}/\sigma_j^2+\Sigma_j^{-1}\right)^{-1}X_{ij}^T\mathbf{y}_{ij}\right), \tag{S1}$$

$$Pr(\gamma_{ij}=0|\mathbf{y}_{ij},X_{ij},\sigma_j^2,\Sigma_j,\pi_{ij})\propto(1-\pi_{ij}), \tag{S2}$$

$$(\gamma_{ij}|\mathbf{y}_{ij},X_{ij},\sigma_j^2,\Sigma_j,\pi_{ij})\sim Bernoulli\left(\frac{(S1)}{(S1)+(S2)}\right)$$

where $\mathbf{y}_{ij}=[y_{ij,1},\ldots,y_{ij,K_{ij}}]^T$.

(f) Update $\boldsymbol{\beta}_{ij}$ from the conditional normal distribution only if $\gamma_{ij}=1$

$$(\boldsymbol{\beta}_{ij}|\gamma_{ij}=1,\mathbf{y}_{ij},X_{ij},\sigma_j^2,\Sigma_j)$$
$$\sim N_p\left(\left(\Sigma_j^{-1}+X_{ij}^T X_{ij}/\sigma_j^2\right)^{-1}X_{ij}^T\mathbf{y}_{ij}/\sigma_j^2,\left(\Sigma_j^{-1}+X_{ij}^T X_{ij}/\sigma_j^2\right)^{-1}\right).$$

(g) Update Σ_j from

$$\left(\Sigma_j^{-1}|\boldsymbol{\beta}_{1j},\ldots,\boldsymbol{\beta}_{m_j,j}\right)\sim Wish\left(a+m_j,\left(R+\sum_{i=1}^{m_j}\boldsymbol{\beta}_{ij}\boldsymbol{\beta}_{ij}^T\right)^{-1}\right).$$

Assay endpoint-specific variance

(h) Update σ_j^2 from

$$(1/\sigma_j^2|\mathbf{y}_{ij},\gamma_{ij},X_{ij},\boldsymbol{\beta}_{ij}\forall i=1,\ldots,m_j)$$
$$\sim Gamma\left(\frac{\nu_0+\sum_{i=1}^{m_j}K_{ij}}{2},\frac{\nu_0\sigma_0^2+\sum_{i=1}^{m_j}\sum_{k=1}^{K_{ij}}(y_{ijk}-\gamma_{ij}(\mathbf{x}_{ijk}^B)^T\boldsymbol{\beta}_{ij})^2}{2}\right).$$