SUPPLEMENT TO "BAYESIAN JOINT MODELING OF CHEMICAL STRUCTURE AND DOSE RESPONSE CURVES"

By Kelly R. Moran, David Dunson, Matthew W. Wheeler, and Amy H. Herring

S.1. Gibbs sampler. To initialize the sampler normally distributed mean 0 variables, constrained to be positive when necessary, can be sampled for all parameters (the authors have also found that choosing a small variance helps keep the sampler from moving in unstable directions). Alternatively, the SVD of $[Y, X]$ can be used to initialize $[\Lambda,\Theta]$, η can be set to $\Lambda^T Y$, and Ξ $X - \Theta\eta$, and the noise variance terms can be initialized based on the residuals when using these initial values.

Assume for notational convenience that the data have been mean-centered prior to analysis (i.e., that $\boldsymbol{\mu}^x$ and $\boldsymbol{\mu}^y$ are fixed to 0-vectors). Let O^Y denote the $D \times N$ matrix with the (d, i)th entry giving the number of observations at dose d for chemical i. Throughout let $\{V\}^{-a}$ denote the set of elements of vector V excepting the a-th entry. Assuming initial values as specified above, the sampler proceeds as follows:

Step 1. **Sample Y-specific factor loading matrix Λ and associated hyper parameters.**

Sample columns of Λ one at a time, letting $\boldsymbol{\lambda}_k$ denote the kth column of Λ. For $k \in 1, \ldots, K$:

- Set Y_k^* to be an $D \times N$ matrix with ith column $y_i^* = (\boldsymbol{y}_i - \sum_{h \neq k} \lambda_h \eta_{h,i})/\eta_{k,i}$. Define S_k^* to be a length-D vector with the dth entry being $\frac{\sum_{i=1}^N y_{d,i}^* \eta_{k,i}^2 O_{d,i}^Y/\sigma_Y^2}{\sum_{i=1}^N \eta_{k,i}^2 O_{d,i}^Y/\sigma_Y^2}$. In other words, S_k^* is an inverse variance weighted mean of Y_k^*.
- Let C_k denote the $D \times D$ GP covariance matrix at all unique dose values, formally defined by the kernel $c_k(d, d') = \alpha_k^2 e^{-\frac{(d-d')^2}{2\ell^2}}$. Let C_k^* be analagous to C_k but with the additive inverse variance component along the diagonal, i.e. $c_k^*(d, d') = c_k(d, d') + 1_{d=d'} \frac{1}{\sum_{i=1}^N \eta_{k,i}^2 O_{d,i}^Y/\sigma_Y^2}$.
- Sample $\boldsymbol{\lambda}_k | \{\lambda_h\}_{h \neq k}, Y, \eta, \alpha_k, \ell, \sigma_Y^2 \sim \mathrm{N}(C_k (C_k^*)^{-1} S_k^*, C_k - C_k (C_k^*)^{-1} C_k)$.

Sample hyper parameters associated with Λ, including the shared $\{\delta_k\}$ that are also used by Θ:

- Let $C_k^{-\alpha}$ denote the $D \times D$ covariance matrix at all unique dose values without the function variance term, formally defined by the kernel $c_k^{-\alpha}(d, d') = e^{-\frac{(d-d')^2}{2\ell^2}}$.
- Assume $l \in L$, where L is a discrete set of possible length-scale values. Sample $\Pr(\ell = l \mid -) = c^* \prod_{k=1}^K \{\det(C_k^l)\}^{-0.5} \exp\{-0.5\ \tau_k^{-1} \boldsymbol{\lambda}_k^T (C_k^l)^{-1} \boldsymbol{\lambda}_k\}$, where C_k^l is defined by covariance kernel $c_k^l(d, d') = e^{-\frac{(d-d')^2}{2l^2}}$ and $c^* = \sum_{l' \neq l} \Pr(\ell = l' \mid -)$ is the normalizing constant.
- Sample $\phi \sim \mathrm{Ga}(\frac{g_\phi + DK}{2}, \frac{\sum_{k=1}^K [\tau_k^{-1} \boldsymbol{\lambda}_k^T (C_k^{-\alpha})^{-1} \boldsymbol{\lambda}_k]}{2})$.
- Define $\tau_k^{(-h)} = \prod_{t=1, t \neq h}^k \delta_t$ for $h = 1, \ldots, K$.

– Sample

$$\delta_1 \mid \text{all} \sim \text{Ga}(a_1 + K(D+S)/2, 1 + \frac{1}{2}\sum_{k=1}^{K}[\tau_k^{(-1)}\phi\boldsymbol{\lambda}_k^T(C_k^{-\alpha})^{-1}\boldsymbol{\lambda}_k + \tau_k^{(-1)}\beta^{-2}\sum_{s=1}^{S}\gamma_{s,k}^{-2}\Theta_{s,k}^2]),$$

$$\delta_h \mid \text{all} \sim \text{Ga}(a_2 + (K-h)(D+S)/2, 1 + \frac{1}{2}\sum_{k=1}^{K}[\tau_k^{(-h)}\phi\boldsymbol{\lambda}_k^T(C_k^{-\alpha})^{-1}\boldsymbol{\lambda}_k + \tau_k^{(-h)}\beta^{-2}\sum_{s=1}^{S}\gamma_{s,k}^{-2}\Theta_{s,k}^2]),$$

$$h = 2, \ldots, K.$$

Finally, set $\tau_k = \prod_{h=1}^{k}\delta_h$ and $\alpha_k^2 = (\phi\tau_k)^{-1}$ for $k = 1, \ldots, K$.

Note that the posterior over ℓ stems from a uniform prior over a discrete grid of possible length scale values. The authors have found a grid size of 100 ranging over "reasonable" length-scale values (i.e., those not implying an effective range smaller than the difference between the closest two dose value, or larger than the range of the data) to be sufficient in the simulation and application described in the paper. However, for larger D or grid sizes it may be computationally preferable to use a Metropolis Hastings step here instead.

Step 2. **Sample latent variable Z corresponding to any non-continuous entries of X.**

Let E^X be the $S \times N$ matrix of expected values of X, i.e. $E^X = \Theta\eta + \Xi\nu$. Then, for $s = 1, \ldots, S$ and $i =, \ldots, N$ such that $x_{s,i}$ is not continuous sample $z_{s,i}$ as follows:

– For binary $x_{s,i}$, sample

$$z_{s,i} \mid E_{s,i}^X \sim \begin{cases} \text{N}_+(E_{s,i}^X, 1), & \text{if } x_{s,i} = 1 \\ \text{N}_-(E_{s,i}^X, 1), & \text{if } x_{s,i} = 0 \end{cases}$$

– For count $x_{s,i}$, sample

$$z_{s,i} \mid E_{s,i}^X \sim \begin{cases} \text{N}_{[t-1,t)}(E_{s,i}^X, 1), & \text{if } x_{s,i} = t \\ \text{N}_{(-\infty,0]}(E_{s,i}^X, 1), & \text{if } x_{s,i} = 0. \end{cases}$$

Note categorical variables should have already been pre-transformed into multiple binary indicators prior to running the Gibbs sampler.

Step 3. **Sample X-specific toxicity-irrelevant components, including factor loadings matrix Ξ, scores ν, and associated hyper parameters.** Define $D = X - \Theta\eta$ to be the $S \times N$ 'residual' toxicity-irrelevant feature matrix.

First sample the $J \times N$ matrix ν, i.e. the matrix of $X-$specific toxicity-irrelevant factors, by column:

– Define $S \times J$ matrix $\Xi^* = \begin{bmatrix} \sigma_1^{-2}\Xi_{\text{ row 1}} \\ \cdots \\ \sigma_S^{-2}\Xi_{\text{ row S}} \end{bmatrix}$, then set $J \times J$ matrix $R^* = (\Xi^T\Xi^* + \text{diag}(1, \ldots, 1))^{-1}$.

– For $i = 1, \ldots, N$, sample $\boldsymbol{\nu}_i \mid \text{all} \sim \text{N}(R^*(\Xi^*)^T D_i, R^*)$.

Next sample the $S \times J$ matrix Ξ, i.e. the matrix of $X-$specific toxicity-irrelevant factor loadings, by row. For $s = 1, \ldots, S$:

– Define $J \times J$ matrix $R_s^* = (\sigma_{X,s}^{-2}\nu\nu^T + \text{diag}(\kappa_{s,1}\omega_1, \ldots, \kappa_{s,J}\omega_J))^{-1}$.

– Sample $\Xi_{\text{ row } s} \mid \text{all} \sim \text{N}(R_s^* \nu (D_{\text{ row } s})^T / \sigma_{X,s}^2, R_s^*)$.

Next sample the local and column-specific shrinkage parameters $\{\kappa_{s,j}\}$ and $\{\zeta_j\}$:

– For $s = 1, \ldots, S$ and $j = 1, \ldots, J$, sample $\kappa_{s,j} \mid \xi_{s,j}, \omega_j, g_\kappa \sim \text{Ga}(\frac{g_\kappa+1}{2}, \frac{g_\kappa+\xi_{s,j}^2 \omega_j}{2})$.

– Define $\omega_j^{(-h)} = \prod_{t=1, t\neq h}^{j} \zeta_t$ for $h = 1, \ldots, J$.

– Sample

$$\zeta_1 \mid \text{all} \sim \text{Ga}(a_1 + JS/2, 1 + \frac{1}{2} \sum_{j=1}^{J} \omega_j^{(-1)} \sum_{s=1}^{S} \kappa_{s,j}^{-2} \Xi_{s,j}^2),$$

$$\zeta_h \mid \text{all} \sim \text{Ga}(a_2 + (J-h)S/2, 1 + \frac{1}{2} \sum_{j=1}^{J} \omega_j^{(-h)} \sum_{s=1}^{S} \kappa_{s,j}^{-2} \Xi_{s,j}^2]),$$

$$h = 2, \ldots, J.$$

Finally set $\omega_j = \prod_{h=1}^{j} \zeta_j$.

Step 4. **Sample X-specific toxicity-relevant components, including factor loadings matrix Θ and its associated shrinkage hyper parameters.** Define $D = X - \Xi\nu$ to be the $S \times N$ toxicity-relevant feature matrix, i.e. the data matrix with the 'residual' toxicity-irrelevant features removed.

First sample the $S \times K$ matrix Θ, i.e. the matrix of $X-$specific toxicity-relevant factor loadings, by row. For $s = 1, \ldots, S$:

– Define $K \times K$ matrix $R_s^* = (\sigma_{X,s}^{-2} \eta\eta^T + \text{diag}(\beta^{-2}\gamma_{s,1}^{-2}\tau_1, \ldots, \beta^{-2}\gamma_{s,K}^{-2}\tau_K))^{-1}$.

– Sample $\Theta_{\text{ row } s} \mid \text{all} \sim \text{N}(R_s^* \eta (D_{\text{ row } s})^T / \sigma_{X,s}^2, R_s^*)$.

Next sample the global and local shrinkage parameters $beta^2$ and $\{\gamma_{s,k}\}$ (note that the shared column-specific shrinkage parameter δ_k was already sampled in the update step for components of Λ):

– Sample $\beta^2 \mid \{\theta_{s,k}\}, \{\tau_k\}, \{\gamma_{s,k}^2\} \sim \text{Ga}(\frac{SK+1}{2}, \frac{1}{t} + \frac{\sum_{k=1}^{K} \tau_k \sum_{s=1}^{S} \theta_{s,k}^2 / \gamma_{s,k}^2}{2})$.

– Sample $\gamma_{s,k}^2 \mid \theta_{s,k}, \tau_k, \beta^2 \sim \text{Ga}(1, \frac{1}{b_{s,k}} + \frac{\tau_k \theta_{s,k}^2}{2\beta^2})$ for $s = 1, \ldots, S$ and $k = 1, \ldots, K$.

Finally sample $\{b_{s,k}\}$ and t, the hyperparameters from the horseshoe prior imposed on elements of Θ :

– Sample $t \mid \beta^2 \sim \text{Ga}(1, 1 + \frac{1}{\beta^2})$.

– Sample $b_{s,k} \mid \gamma_{s,k}^2 \sim \text{Ga}(1, 1 + \frac{1}{\gamma_{s,k}^2})$ for $s = 1, \ldots, S$ and $k = 1, \ldots, K$.

Step 5. **Sample shared toxicity-relevant factor matrix η.** Define concatenated $(D+S) \times K$ loadings matrix $\Omega = \begin{bmatrix} \Lambda \\ \Theta \end{bmatrix}$ and $(D+S) \times N$ data matrix $W = \begin{bmatrix} Y^{\text{sum}} \\ X - \Xi\nu \end{bmatrix}$ where Y^{sum} is a $D \times N$ matrix with entry $[h, i]$ giving the sum across replicates of the response values for chemical i at dose index h (or left as missing for chemical i with no observations at dose index h). Explicitly, $Y^{\text{sum}}[h, i] = \sum_{r=1}^{R[i]} \boldsymbol{y}_{i[r],h}$. Let $O_{h,i}$ denote the number of observations of chemical i at dose index

h. For example, if chemical i has two replicates observed at dose indices $\{1,3,5\}$ and $\{1,2,5,6\}$, respectively, then (assuming $D=6$) $\boldsymbol{O}_i = [O_{1,i},\ldots,O_{6,i}]'$ equals $[2,1,1,0,2,1]'$.

Sample the $K \times N$ matrix η by column. For $i = 1,\ldots,N$:

- Let $U_i^* = \{d_1^*,\ldots,d_U^*\}$ be the set of $U \leq D$ unique dose values at which chemical i has at least one observation (i.e., $d_u^* \in U_i^*$ iff $O^Y_{d_u^*,i} > 0$). Use $M_{[U_i^*]}$ to denote the matrix with only the relevant doses selected (across either rows or columns, depending on which is the relevant dimension). For example, $\Omega_{[U_i^*]}$ is the $(U+S)\times K$ matrix with the first U rows being the subset of the original first D rows for doses in U_i^*.
- Define $\sigma^2_{Y,i}$ to be a length-D vector with entry $\sigma^2_{Y,i[d]} = \begin{cases} \sigma_Y^2/O^Y_{d,i}, & \text{if } O^Y_{d,i} > 0 \\ \sigma_Y^2, & \text{if } O^Y_{d,i} = 0 \end{cases}$.
- Set Ω_i^* to be the $(D+S)\times K$ matrix with row p being the pth row of Ω times $1/\sigma^2_{Y,i[d]}$ for the first D rows, and the pth row of Ω times $1/\sigma^2_{X,p-D}$ for the remaining S rows.
- Define the $K \times K$ matrix $R^* = (\Omega^T_{[U_i^*]}\Omega^*_{i[U_i^*]} + \text{diag}(1,\ldots,1))^{-1}$.
- Sample $\boldsymbol{\eta_i} \mid \text{all} \sim \text{N}(R^*(\Omega^*_{i[U_i^*]})^T W_{[U_i^*],i}, R^*)$.

Step 6. **Sample entries of noise variance parameters Σ_X and Σ_Y.** Let $E^Y = \Lambda\eta$ and $E^X = \Theta\eta + \Xi\nu$ denote the mean of Y and X, respectively.

First sample σ_Y^2, the common noise variance term for the dose response curves:

- Let D_i^* denote the set of all (not necessarily unique) doses at which chemical i has observations. E.g., D_i^* could be $\{0,0,0.1,0.1,0.2,0.3\}$.
- Define $\text{RSS}_Y = \sum_{i=1}^N \sum_{d\in D_i^*}(Y_{d,i} - E^Y_{d,i})^2$ and $N_Y = \sum_{i=1}^N |D_i^*|$.
- Sample $\sigma_Y^2 \mid \text{all} \sim \text{Ga}(\frac{a_{\sigma_Y}+N_Y}{2}, \frac{b_{\sigma_Y}+\text{RSS}_Y}{2})$.

Next sample $\{\sigma^2_{X,s}\}$, the feature-specific noise variance terms, for $s = 1,\ldots,S$:

- Define $\text{RSS}_{X,s} = \sum_{i=1}^N (X_{s,i} - E^X_{s,i})^2$.
- Sample $\sigma^2_{X,s} \mid \text{all} \sim \text{Ga}(\frac{a_{\sigma_X}+N}{2}, \frac{b_{\sigma_X}+\text{RSS}_{X,s}}{2})$.

Note that an informative prior on σ_Y^2 can utilize the variance of low-dose observations in related assays, i.e. observations for which no activity is expected. Such a prior encourages the model to learn structure in the curves rather than simply learning a large noise variance.

S.2. Additional simulation information and results.

S.2.1. *Sampling details.* In the BS^{3}FA model, the B-FOSR, and the BAABTP model, four chains are run with the initial 20000 samples discarded as burn-in and every 10th draw of the subsequent 20000 samples saved. For the BS^{3}FA model the number of toxicity relevent and irrelevant factors is set to be the true number of factors plus 5, the intention being to mimic the act of providing a conservative 'upper bound' in the real data scenario. For the BAABTP model the number of mixture components is set to 5. For LASSO, the dose levels, all chemical features, and all pairwise interactions are included as covariates.

S.2.2. *Distance performance.* For distance performance, BS^{3}FA is compared to Euclidean distance in the full feature space and in PCA space (i.e., with no supervision). The distance between chemicals in η space is used as the truth when assessing model chemical similarity performance. Correlation between model predicted distance and true distance is used as a scale-invariant measure of model performance. Predicted distance is a metric defined by the latent factors in each method (i.e., by model-predicted η for BS^{3}FA and by the principal component scores in PCA). Coverage is assessed for our model, B-FOSR, BAABTP, and a straw man model in which the standard deviation about low-dose response values is used to approximate a 95% confidence interval around the observed training mean. Note that JIVE was not included as a competitor algorithm because the amount of missingness introduced by omitting the observations of Y for the test data is is too high for the algorithm (the use of the `jive()` function in the **R** package **r.jive** resulted in an error message: "Unable to complete matrix, too much missing data").

S.2.3. *Inputs provided for each model.* The inputs to B-FOSR are all chemical features. The inputs to LASSO are the dose, chemical features, and all pairwise interactions therein. The features used in the BAABTP model are PC_{95}. The LASSO model was trained using the `cv.glmnet` function from the **glmnet** package and predictions were made using the predict function from the resulting object with `s='lambda.1se'`.

S.2.4. *Simulated data structure with truth-model alignment.* For all simulations, data on 300 'chemicals' was created, comprised of dose-response curves and structure information. The number of unique doses was set to $D = 10$, and the number of chemical features was set to $S = 20$. The noise terms were set to be homoscedastic with $\sigma_Y = 0.2$ and $\sigma_X = 0.1$. The true dimension of the latent toxicity-relevant space was varied $K \in \{1, 3, 5\}$, as was that of the latent toxicity-irrelevant space $J \in \{0, 5, 10, 15, 20\}$. In each simulated data set, 25% of the chemical dose-response curves are hidden from the models as test data, while the remaining 75% are used as training data (note that feature data are available for all chemicals, not just training data). At each simulation setting, 100 simulated data sets are created.

The toxicological 'data' were simulated so as to mimic realistic dose response curves. A smooth factor loadings matrix Λ is created in each simulation by first smoothing the real Attagene PXR data using a functional factor model, then randomly sampling 500 smoothed dose response curves from this set and setting Λ to be the first K loadings from the SVD of this smooth subset of curves. A sparse factor loadings matrix Θ is created for each simulation by creating a $K \times S$ standard normal matrix M, then using the `spEigen()` function from the **sparseEigen R** package to compute sparse orthogonal eigenvectors of the covariance matrix of M (i.e., setting Θ equal to `spEigen(t(M) %*% M, q=K, rho=0.2)$vectors`).

S.2.5. *Simulated data structure with misalignment between the model and the true data generating process.* We also perform a simulation study in which the structure assumed by BS^{3}FA is *not* the true data generating process, to mimic the (likely) scenario of a misalignment between our assumptions and the truth. The hope is that BS^{3}FA is still able to perform as well as competitors under such a misalignment.

Define S_{relevant} to be the number of toxicity-relevant features and $S_{\text{irrelevant}}$ to be the number of toxicity-irrelevant features. Then $S = S_{\text{relevant}} + S_{\text{irrelevant}}$ is the total number of features in X. To assess how the model performs when the true data generating process does not match the data generating process assumed by BS^{3}FA, data were simulated assuming a polynomial relationship

with dose. The misalignment simulation sets $y_{i,d} = \sum_{m=1}^{S_{\text{relevant}}} x_{i,m}\ d^m$. That is, the response is a polynomial function of dose with the parameters controlling the shape of the polynomial being the true "toxicity-relevant" entries in $\boldsymbol{x}_i$. Thus the shape of the dose response curves does depend on some features in $\boldsymbol{x}_i$, but not in the way assumed by BS^{3}FA (although we note that when $S_{\text{relevant}} = 1$ the polynomial is linear and can in fact be represented by the BS^{3}FA model). We vary the number of toxicity-relevant features $S_{\text{relevant}} \in \{1, 2, 3\}$ and the number of toxicity-irrelevant features $S_{\text{irrelevant}} \in \{0, 5, 10, 20, 30\}$.

For all simulations and choices of S_{relevant} and $S_{\text{irrelevant}}$, data on 300 'chemicals' was created. The number of unique doses was set to $D = 10$. The noise about the true dose response mean is assumed to be homoscedastic with $\sigma_Y = 0.2$. In each simulated data set, 25% of the chemical dose-response curves are hidden from the models as test data, while the remaining 75% are used as training data (note that feature data are available for all chemicals, not just training data). At each simulation setting, 100 simulated data sets are created.

S.2.6. *Additional results for model-truth-aligned simulation.* Within a plot throughout this section, each point shows the mean of the metric of interest from the 100 simulated data sets having the specified settings. The lower and upper bands around this point give the 2.5 and 97.5 percentiles of the performance values across the 100 simulations, respectively.

Figures S.1 through S.3 show visualizations of MSE, correlation, and distance coverage results for the BS^{3}FA model alone. It is clear that the model does best when the dimension of both $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ are smallest. That is, performance degrades as K and/or J increase. However, this degradation is relatively minimal, and the model performance overall is still quite good even with large K and/or J.

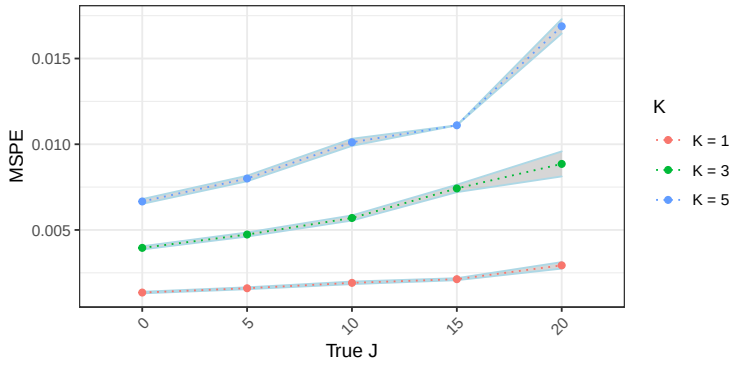

Fig S.1: MSPE between BS^{3}FA-predicted dose-response profile and true dose-response profile for hold-out "chemicals".

Figure S.4 shows the coverage for the hold-out chemicals' dose-response values. The BS^{3}FA model generally has close-to-nominal coverage and is robust to increasing "superfluous" information in X, whereas the BAABTP model is harmed by its presence. This phenomenon is likely due to the distance based kernel and the flexibility of the BAABTP model; the kernel cannot separate relevant from irrelevant features, which leads to overfitting of the training data and poor performance on holdout data.

Table S.1 shows the mean and SD of the test chemicals' 95% simultaneous credible band widths for each model across all simulations. It is preferable to have the narrowest band width subject to (at least) nominal coverage. Combining these band width results with our coverage results, we see that

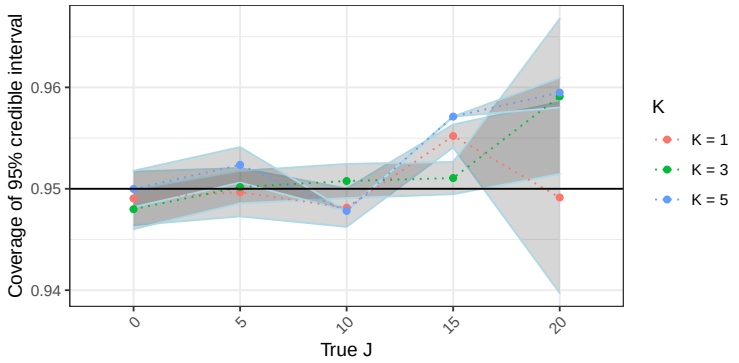

Fig S.2: Proportion of hold-out chemicals' (noisy) data points covered by the BS^{3}FA 95% credible/confidence intervals.

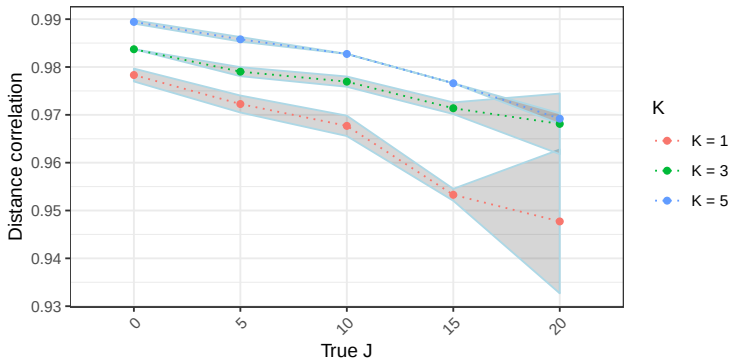

Fig S.3: Correlation between BS^{3}FA-predicted distances in $\boldsymbol{\eta}$ space and true distance in $\boldsymbol{\eta}$ space.

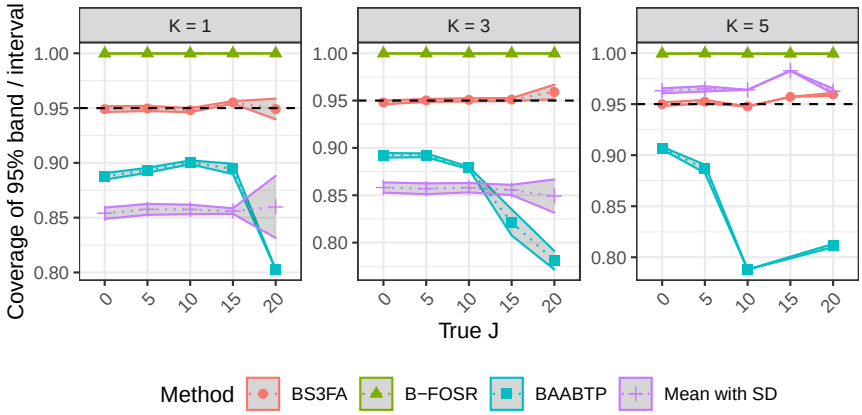

Fig S.4: Proportion of hold-out chemicals' (noisy) data points covered by the 95% credible/confidence intervals for each of the methods. Each subplot shows the result across methods for a given true shared subspace dimension K.

BS^{3}FA has the narrowest simultaneous bands subject to (at least) nominal coverage across values of K and J. B-FOSR consistently over-covers and has wider simultaneous bands than BS^{3}FA, whereas BAABTP consistently under-covers and has narrower simultaneous bands than BS^{3}FA. The straw-man mean with SD estimates have bands that over- or under-cover and band widths that are narrower or wider than BS^{3}FA depending on K and J.

		$J=0$	$J=5$	$J=10$	$J=15$	$J=20$
BS^{3}FA	$K=1$	1.14 (0.03)	1.14 (0.03)	1.14 (0.04)	1.15 (0.04)	1.15 (0.04)
	$K=3$	1.17 (0.04)	1.16 (0.04)	1.19 (0.05)	1.21 (0.06)	1.21 (0.08)
	$K=5$	1.18 (0.04)	1.20 (0.05)	1.22 (0.07)	1.24 (0.08)	1.26 (0.12)
B-FOSR	$K=1$	1.57 (0.07)	1.67 (0.13)	1.69 (0.15)	1.72 (0.18)	1.75 (0.21)
	$K=3$	1.65 (0.15)	1.76 (0.18)	1.87 (0.23)	1.93 (0.26)	2.07 (0.35)
	$K=5$	1.72 (0.22)	1.91 (0.27)	2.08 (0.32)	2.21 (0.34)	2.27 (0.38)
BAABTP	$K=1$	0.69 (0.03)	0.68 (0.02)	0.69 (0.06)	0.80 (0.18)	0.81 (0.15)
	$K=3$	0.72 (0.04)	0.70 (0.03)	0.85 (0.21)	0.91 (0.17)	0.99 (0.22)
	$K=5$	0.74 (0.05)	0.74 (0.04)	0.92 (0.26)	1.07 (0.25)	1.20 (0.26)
Mean with SD	$K=1$	0.78 (0.02)	0.79 (0.03)	0.80 (0.04)	0.78 (0.03)	0.79 (0.03)
	$K=3$	1.19 (0.05)	1.21 (0.08)	1.26 (0.07)	1.21 (0.07)	1.25 (0.07)
	$K=5$	2.75 (0.15)	2.68 (0.16)	2.73 (0.16)	2.67 (0.20)	2.68 (0.20)

Table S.1: Hold-out chemical 95% simultaneous credible band widths for each method yielding interval estimates. Shown is the mean (SD) band width across all chemicals and simulations.

Of course, predictive ability and ability to characterize distance in $\boldsymbol{\eta}$ space are not the only model capabilities of interest. Model components may be interrogated for interpretation as well. Table S.2 provides an assessment of the fit for various model components of BS^{3}FA. The MSE for σ_x^2 is poor when J is small because there is a lack of identifiability between the non-toxicity-relevant structure component $\Xi\boldsymbol{\nu}$ and the noise term σ_x^2. Since the latent dimension J is set to an over-estimate when running BS^{3}FA, there is "room" in $\Xi\boldsymbol{\nu}$ to absorb σ_x^2. As J increases, the model becomes greedier and needs to use up the $\Xi\boldsymbol{\nu}$ component to explain non-noise variability in X not shared with Y. On the other hand, the terms $\Lambda\Lambda'$, $\Theta\Theta'$, and σ_y^2 are all identifiable and are consistently well estimated by our model.

Recall that a chemical is deemed active if its global Bayesian p-value is less than 0.05. Table S.3 shows the true positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR) of the B-FOSR method on the simulated data. Across all values of K and J the TPR is quite high, and the FPR and FDR are middling. Note that the TPR, FPR, and FDR all increase with K and seem fairly invariant to changes in J. That is, the model loses specificity when the dimension of the latent toxicity-relevant space increases.

S.2.7. *MSPE, coverage, and distance results for model-truth-misaligned simulation.* Figure S.5 shows the mean squared predictive error (MSPE) for the hold-out chemicals' dose-response mean functions when there is misalignment between the structure assumed by the BS^{3}FA model and the true data generating process. In spite of this misalignment, BS^{3}FA is able to predict similarly to or better than the competitors. As with the well-aligned simulation, BS^{3}FA appears robust to increasing "superfluous" information in X.

A similar story can be seen in the coverage and distance results (Figures S.6 and S.7) for the

		$J = 0$	$J = 5$	$J = 10$	$J = 15$	$J = 20$
$\Lambda\Lambda'$	$K = 1$	4e-05 (3e-05)	4e-05 (2e-05)	4e-05 (2e-05)	4e-05 (2e-05)	4e-05 (2e-05)
	$K = 3$	3e-05 (0)	5e-05 (1e-05)	7e-05 (0)	8e-05 (4e-05)	0.00011 (5e-05)
	$K = 5$	4e-05 (2e-05)	5e-05 (2e-05)	5e-05 (0)	3e-05 (0)	8e-05 (3e-05)
$\Theta\Theta'$	$K = 1$	9e-05 (3e-05)	0.00011 (6e-05)	1e-04 (6e-05)	7e-05 (3e-05)	0.00015 (0.00011)
	$K = 3$	3e-05 (0)	3e-05 (0)	4e-05 (0)	4e-05 (1e-05)	4e-05 (2e-05)
	$K = 5$	2e-05 (0)	2e-05 (0)	1e-05 (0)	2e-05 (0)	3e-05 (1e-05)
σ_y^2	$K = 1$	0.00011 (0.00015)	0.00011 (0.00017)	2e-05 (2e-05)	1e-05 (1e-05)	1e-05 (2e-05)
	$K = 3$	1e-05 (0)	2e-05 (0)	5e-05 (0)	3e-05 (3e-05)	3e-05 (4e-05)
	$K = 5$	3e-05 (3e-05)	3e-05 (3e-05)	1e-05 (0)	3e-05 (1e-05)	8e-05 (7e-05)
σ_x^2	$K = 1$	0.4350 (0.0373)	0.0369 (0.0052)	0.0121 (0.0014)	0.0064 (0.0002)	0.0045 (0.0003)
	$K = 3$	0.2065 (0.0267)	0.0257 (0.0029)	0.0103 (0.0009)	0.0064 (0.0003)	0.0051 (0.0001)
	$K = 5$	0.1123 (0.0182)	0.0216 (0.0029)	0.0091 (0.0008)	0.0056 (0.0003)	0.0060 (0.0001)

Table S.2: MSE between the true and estimated values for each model component. These components characterize the structured toxicity-relevant directions of variation and the noise variance in the simulated data.

		$J = 0$	$J = 5$	$J = 10$	$J = 15$	$J = 20$
TPR	$K = 1$	0.82 (0.07)	0.82 (0.07)	0.81 (0.08)	0.82 (0.04)	0.76 (0.07)
	$K = 3$	1.00 (0.01)	1.00 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)
	$K = 5$	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
FPR	$K = 1$	0.13 (0.09)	0.21 (0.12)	0.21 (0.09)	0.11 (0.09)	0.18 (0.09)
	$K = 3$	0.31 (0.10)	0.38 (0.10)	0.36 (0.09)	0.38 (0.10)	0.39 (0.11)
	$K = 5$	0.45 (0.10)	0.48 (0.13)	0.49 (0.09)	0.48 (0.09)	0.49 (0.08)
FDR	$K = 1$	0.13 (0.07)	0.19 (0.09)	0.20 (0.08)	0.11 (0.08)	0.18 (0.09)
	$K = 3$	0.24 (0.06)	0.28 (0.07)	0.26 (0.06)	0.27 (0.07)	0.28 (0.08)
	$K = 5$	0.30 (0.07)	0.31 (0.07)	0.33 (0.06)	0.32 (0.06)	0.33 (0.05)

Table S.3: True positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR) for the B-FOSR model under the proposed method of assessing whether a chemical is active. A perfect classifier has a TPR of 1 and an FPR/FDR of 0.

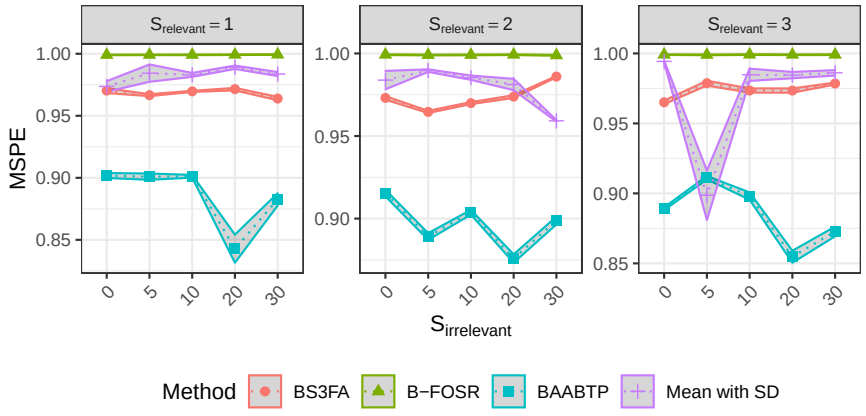

Fig S.5: Mean squared predictive error (MSPE) for the hold-out chemicals' dose-response mean functions under the polynomial model.

misaligned simulation as for the model-truth-aligned simulation. Namely, BS^{3}FA has closest-to-nominal (albeit still somewhat high) coverage, while other methods are less well calibrated. BS^{3}FA has very stable, high correlation between the true Euclidean distance matrix for the relevant parts of X, and that of the model-predicted η. Direct PCA, FOSR-VS, and Euclidan distance suffer immediately and harshly when $S_{\text{irrelevant}}$ exceeds 0.

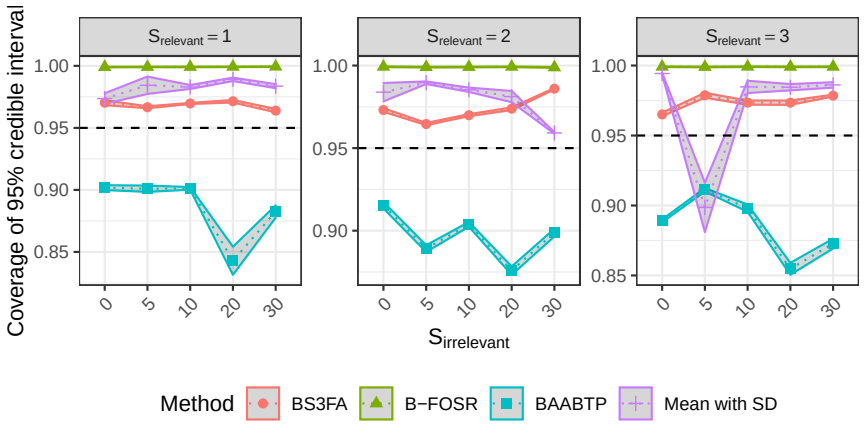

Fig S.6: Proportion of hold-out chemicals' (noisy) data points covered by the 95% credible/confidence intervals for each of the methods under misalignment between the true and BS^{3}FA-assumed structure.

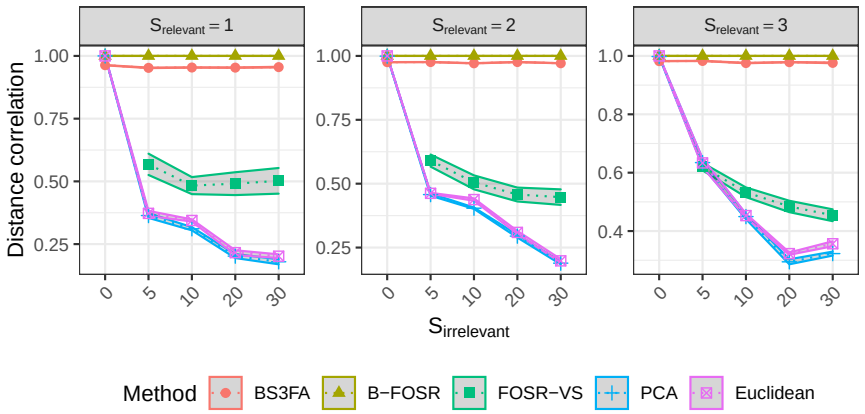

Fig S.7: Correlation between entries in the true pairwise distance matrix (i.e., the Euclidean distance between true relevant dimensions of X) and the predicted pairwise distance matrix for holdout chemicals under misalignment between the true and BS^{3}FA-assumed structure.

S.3. Additional ToxCast run information.

S.3.1. *Data download.* A pre-cleaned **R** data file holding a subset of the information from the ToxCast database was created by Dr. Matthew Wheeler and can be downloaded from `https://1drv.ms/u/s!AoYFThhStiORt0YjFGEQDKjBa4BZ`. The download, `gain.Rdata`, is a file containing the dose-response information for all Phase 1, Phase 2, and e1k chemicals tested across all ToxCast assays. For this analysis, only results for the the Attagene PXR assay (i.e., the assay having value 135 for the variable `aeid`) were saved and are provided with this manuscript as the file `atg_pxr_data.Rdata`. The full ToxCast data are available for download from `https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data`.

S.3.2. *Results and diagnostics.* We consider K to be chosen "large enough" if the smallest value of $1/\tau_j$ is below 0.01. Similarly, we consider J to be chosen "large enough" if the smallest value of $1/\omega_j$ is below 0.01. Figure S.8 shows the ordered values $1/\tau_j$ and $1/\omega_j$ for the Tox run.

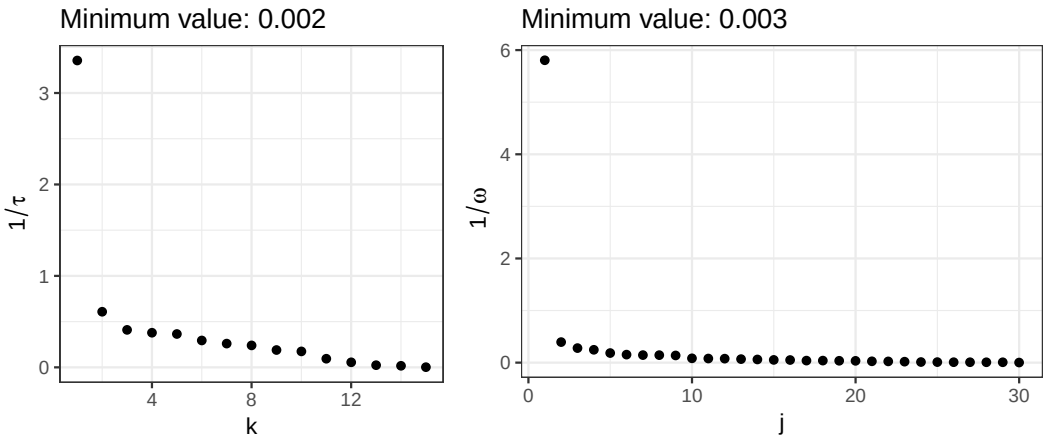

Fig S.8: Ordered variance elements driving column-specific shrinkage of the factor loadings matrices Λ and Θ (left), and Ξ (right).

Figure S.9 shows posterior predictive plots for the maximum observed response value for select test chemicals with the actual maximum observed response value denoted by a dashed vertical line. The chemicals shown in the first two rows are the same hold-out chemicals as those shown in Figure 19 and the first row of Figure 20 in the main paper; the final row shows three chemicals having particularly high MSPE (i.e. poor predictions). Unsurprisingly, the poorer the overall fit of the predicted dose-response profile to the data, the farther toward the boundary of the posterior predictive distribution the observed data are. For example, the bottom left subplot, in which model posterior predictions are all much lower than the observation, reflects the fact that this chemical was chronically under-predicted by the model. When the predicted dose response profile seems well aligned with the data the posterior predictive plots show minimal signs of misalignment between the posterior and the data.

Figure S.10 shows the results for hold-out chemicals having poorly predicted dose response curves. These poor predictions may be a sign that the Mold2 structure information doesn't contain enough toxicity-relevant information to fully inform the dose response curve shapes, that we do not have enough training data, and/or that underlying model assumptions are incomplete. Further work should explore which chemicals are poorly predicted and why.

Trace plots for the predicted dose response profiles of hold-out chemicals show good mixing (a randomly sampled set of chemicals are shown across multiple doses in Figure S.11), as do the noise variance terms for the data (samples of σ_Y are shown in Figure S.12).

While there are some inconsistencies between predicted values across multiple chains at higher doses (see Figure S.13), the general curve shapes and predicted summary quantities are consistent across chains (see Figure S.14).

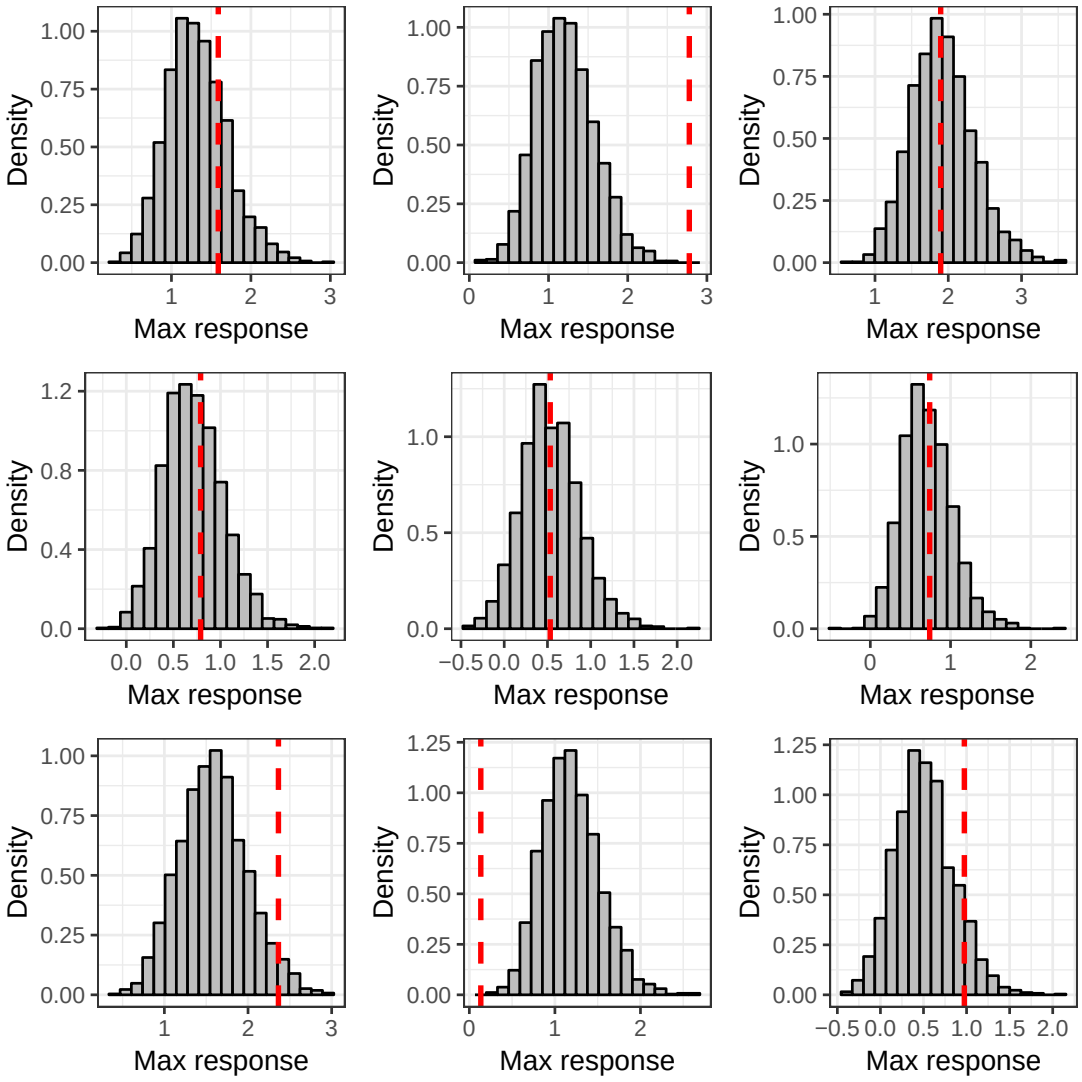

Fig S.9: Posterior predictive plots for the maximum observed response value for select test chemicals. The true maximum observed response value is indicated by a dashed vertical line. The posterior predictive does not appear misaligned with the data except in the case of poorly fitting curves generally or apparent data outliers. For example, the bottom left plot corresponds to general model under prediction of the dose response curve for that chemical.

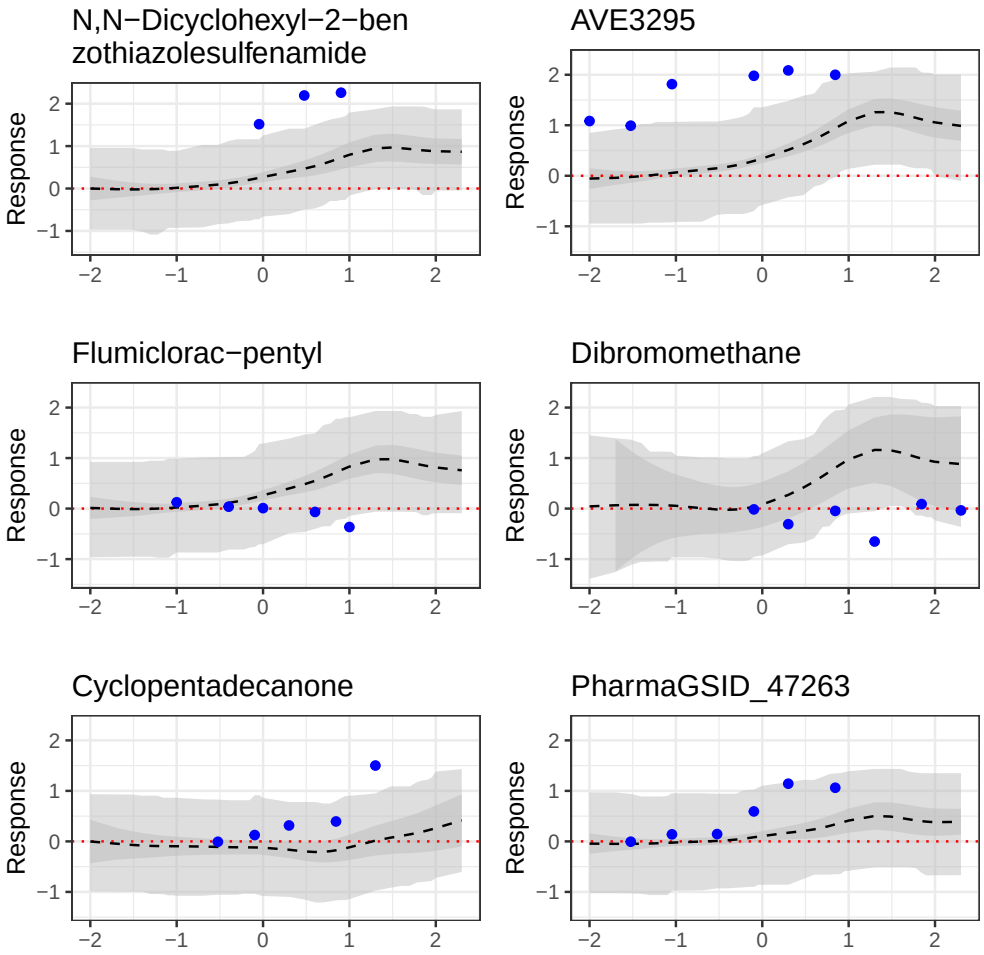

Fig S.10: Results for select hold-out chemicals for which the fit between predicted dose response curve and observed data is particularly poor. Shown are predicted average dose-response curve (dashed black line), 95% simultaneous interval for expected dose-response curve (darker grey ribbon), and 95% credible interval for observed data (lighter grey ribbon). Data (held out in training) are solid blue points. Top: both chemicals have abnormally high response values. While BS^{3}FA predicts both to be activating, it does not predict the height of the activity. Middle: these chemicals both appear non-activating, but the model predicted that these chemicals were activating. Bottom: both chemicals appear activating, but the model predicted that these chemicals were non-activating or of low activity.

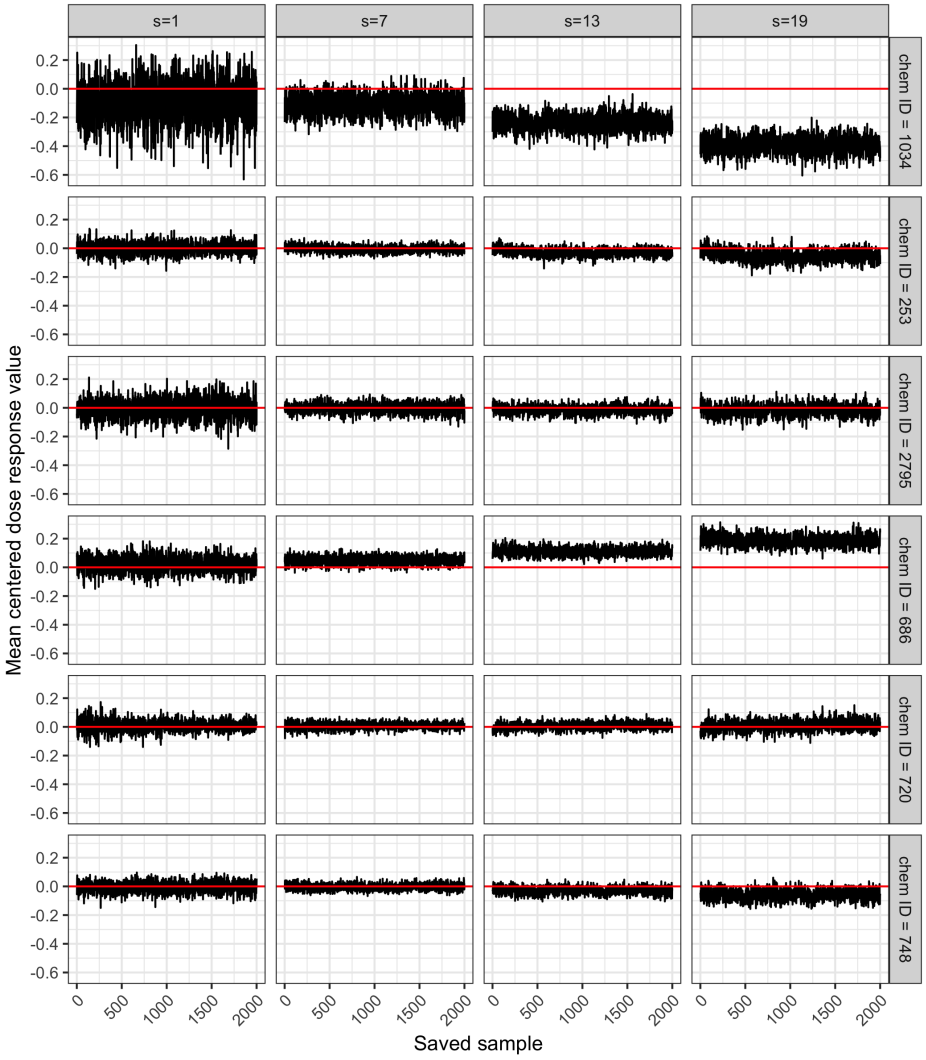

Fig S.11: Trace plots for randomly selected hold-out chemicals' response at multiple doses (indexed by s in column headers).

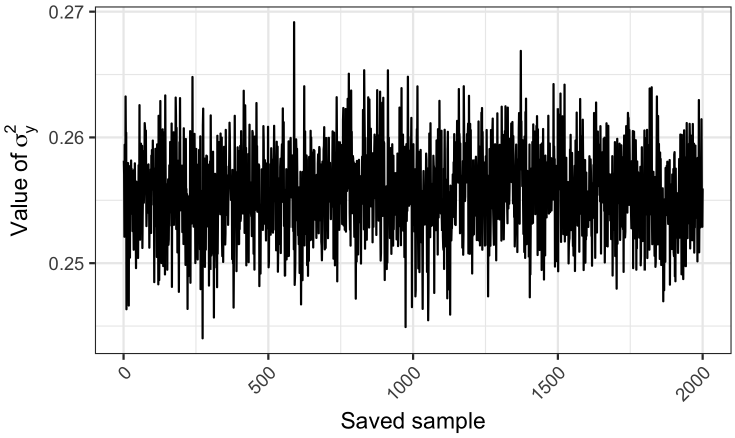

Fig S.12: Trace plot showing samples of the noise variance term for Y.

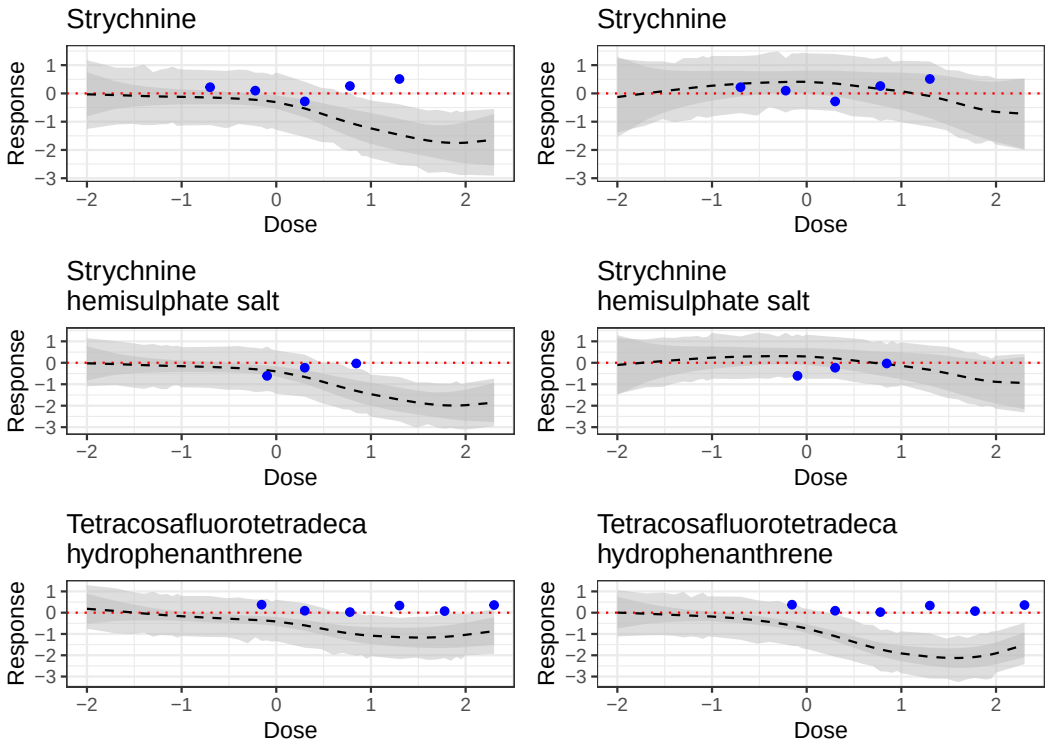

Fig S.13: The predicted dose response curves for hold-out chemicals having the largest divergence between predicted mean across two chains. While the exact shape of the predicted curve differs (e.g., note how in the bottom row chain 1 predicts a slightly flatter curve with less of a U shape at the end than chain 2), small regions of multi-modality do not concern us. Even in the "worst" behaving chemicals the general direction of effect is similar, and the chemical profiles predicted are consistent for the large majority of chemicals.

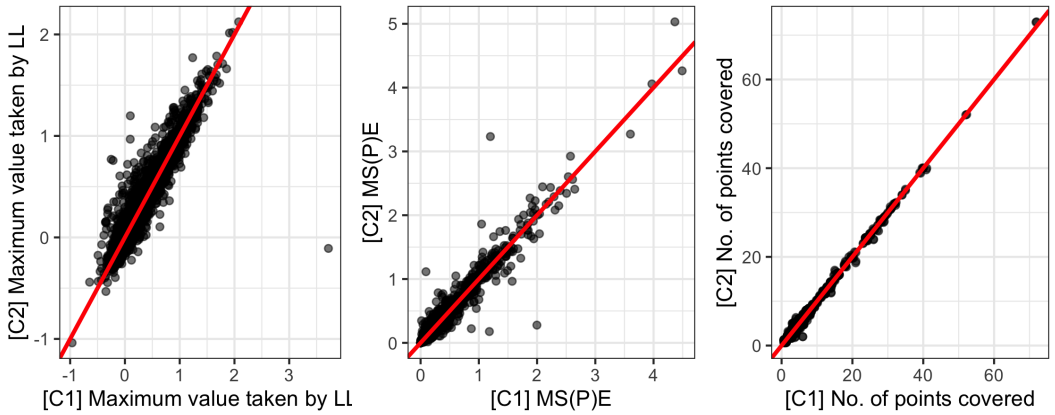

Fig S.14: A comparison between two chains of the following chemical-specific value. From left to right: the maximum value of the lower bound for the dose response curve; the mean square (predictive) error for training (test) chemicals; the number of points covered by the model 95% posterior data credible interval interval. Red line denotes $x = y$.

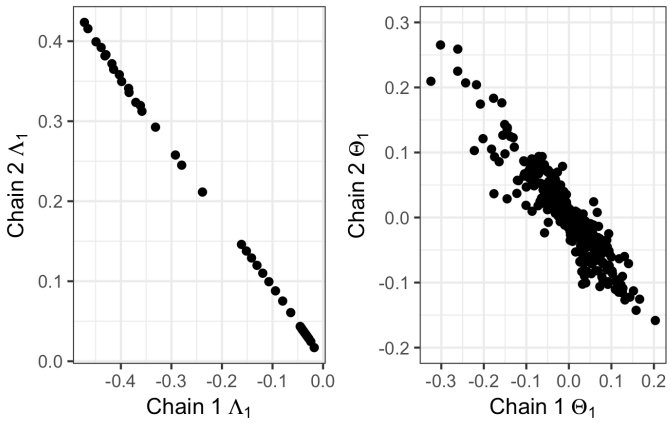

Fig S.15: A comparison between two chains of model predicted components. Left: the column of Λ having the largest 2-norm. Right: the associated column of Θ. Note that scales and signs differ across chains, but general relationships remain consistent.

Figures S.16 through S.18 show posterior predictive histograms for randomly selected chemical feature sets X. In general, the model posterior predictions are consistent with observed data. An exception to this is for a small number of continuous features having heavier-than-normal tails, for which the model posterior predicted medians are consistent with observed data but model posterior predicted maximum absolute values are less than the true maximum absolute value. Exploring allowing for heavier tails in the feature data is an avenue of future research.

S.4. Data structure for chemical features. The Mold2 software was downloaded from `https://www.fda.gov/scienceresearch/bioinformaticstools/mold2/default.htm`. A description of the process for generating Mold2 descriptors using the information provided by ToxCast is provided with this manuscript as the file `workflow.txt`.

S.4.1. *"Identical" chemicals.* Our empirical examination showed that the majority of chemicals grouped together via having the same-Mold2-output the dose-response profiles are similar up to noise (see some example output in Figure S.20, in which the grey/black points are chemicals having different SMILES but identical Mold2 output). We suspect that augmenting Mold2 is potentially useful, but not critical for the purpose of the illustrative application in the main paper.

S.4.2. *Count features.* The model can accommodate count data via the underlying normal assumption and rounding operator described in the main paper. However, this rounding operator is more computationally expensive than simply treating a variable as continuous, and a log transformation in many cases will allow a count feature to well-approximate a continuous normal variable.

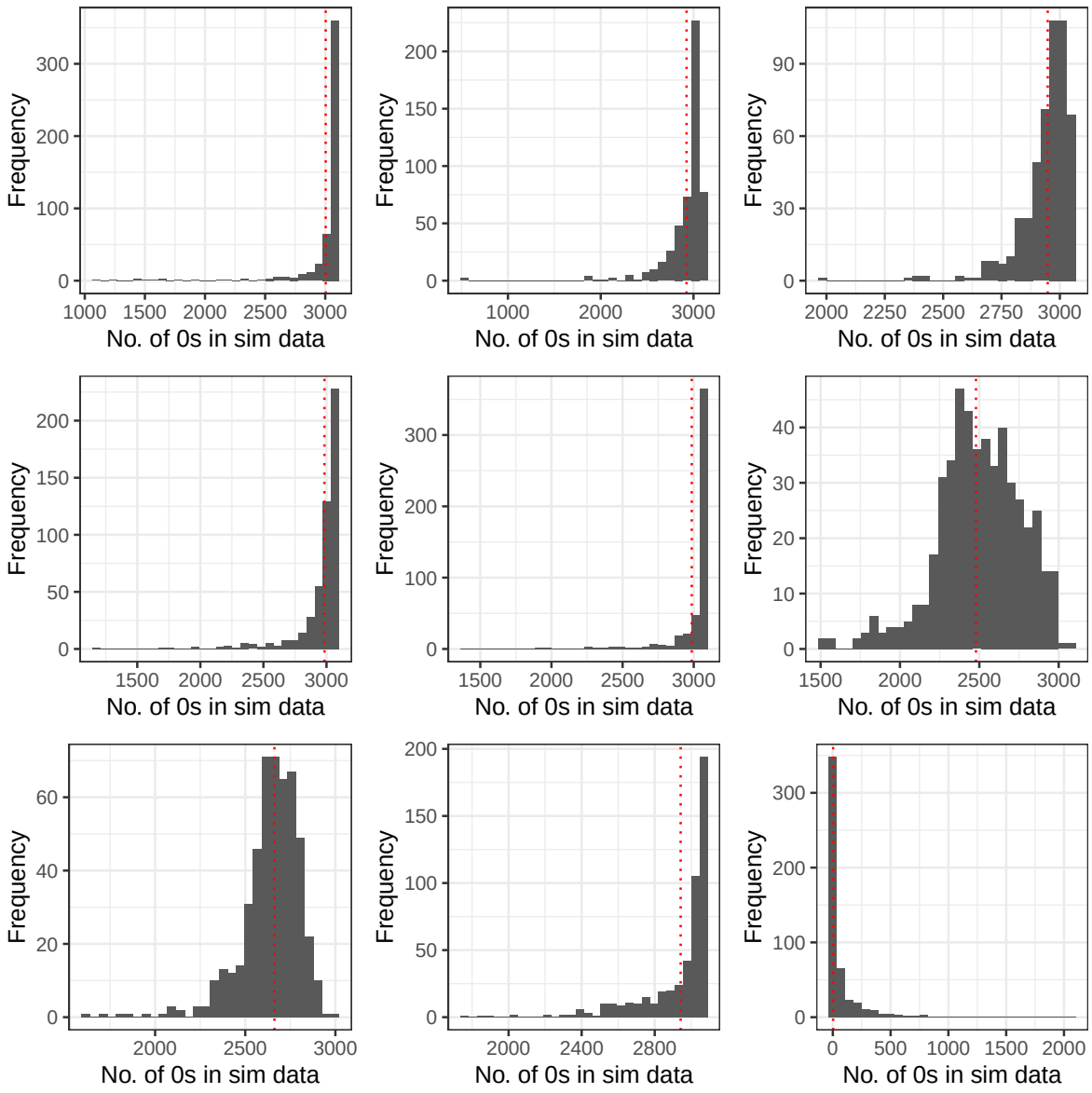

Fig S.16: Posterior predictive histograms for randomly selected count chemical features showing model-predicted draws of the number of 0s in simulated datasets (histogram) and the observed number of 0s in the real dataset (vertical red line).

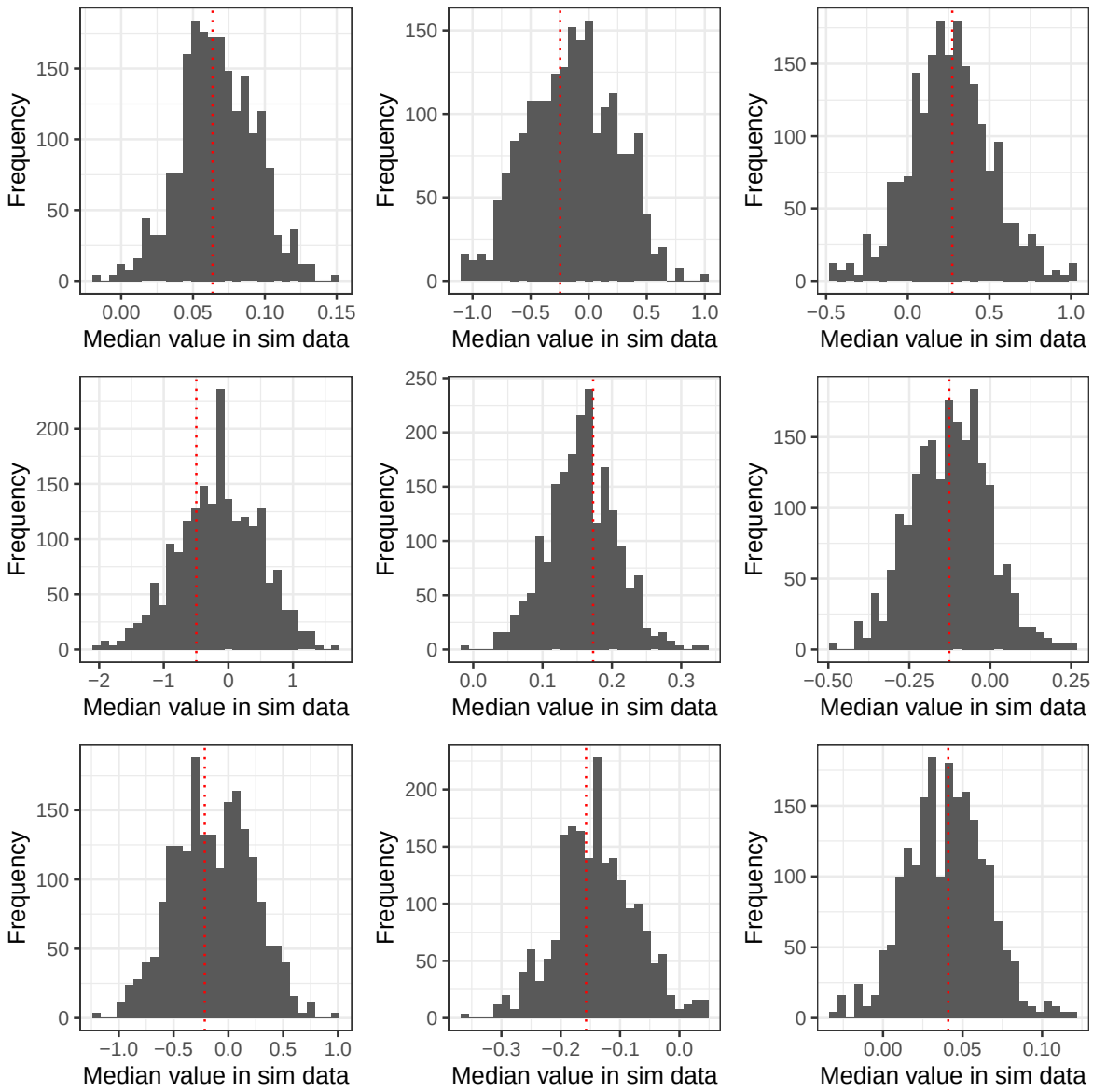

Fig S.17: Posterior predictive histograms for randomly selected continuous chemical features showing model-predicted draws of the median value in simulated datasets (histogram) and the observed median in the real dataset (vertical red line).

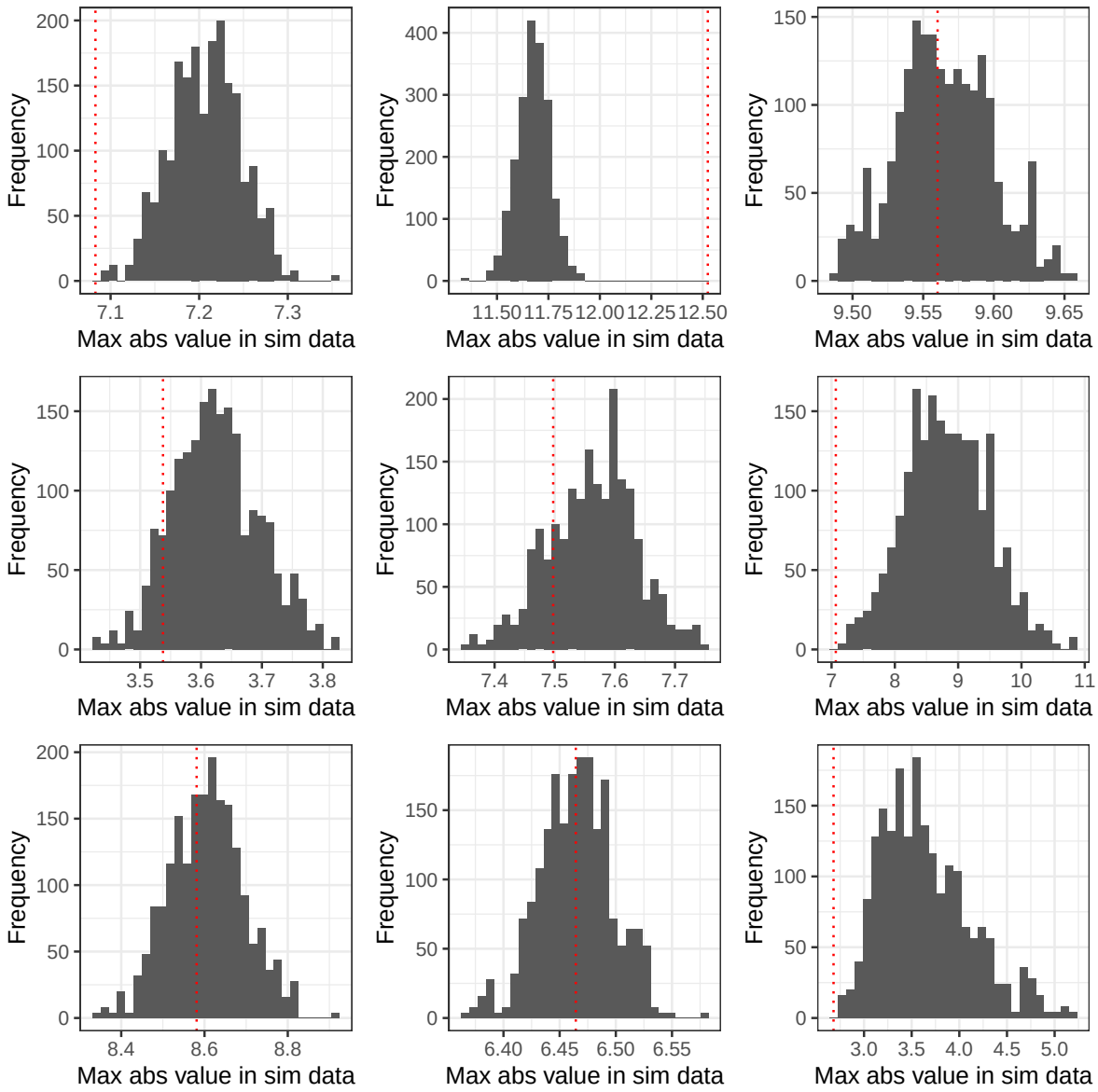

Fig S.18: Posterior predictive histograms for randomly selected continuous chemical features showing model-predicted draws of the maximum absolute value in simulated datasets (histogram) and the observed maximum absolute value in the real dataset (vertical red line).

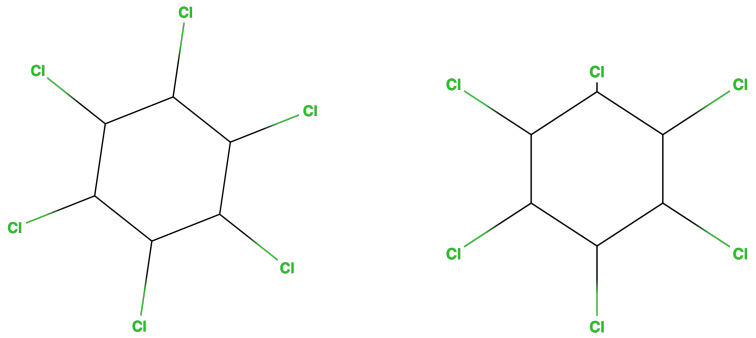

Fig S.19: An example of a pair of chemicals having identical Mold descriptor sets. On the left is beta-Hexachlorocyclohexane and on the right is delta-Hexachlorocyclohexane. The chemical SMILES are:
Cl[C@H]1[C@H](Cl)[C@@H](Cl)[C@H](Cl)[C@@H](Cl)[C@@H]1Cl and
Cl[C@H]1[C@H](Cl)[C@@H](Cl)[C@H](Cl)[C@H](Cl)[C@@H]1Cl, respectively.

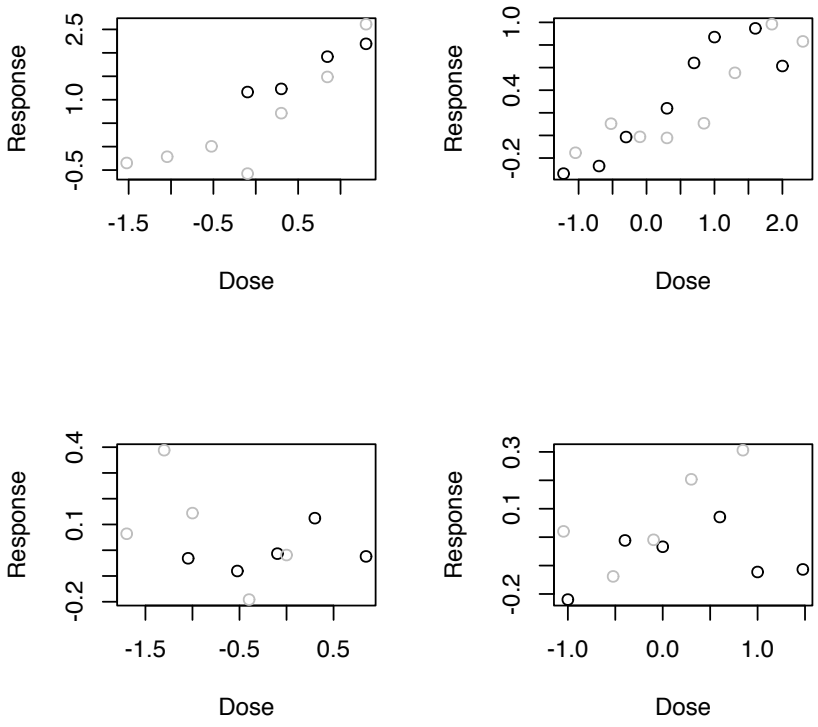

Fig S.20: Example dose response profiles for pairs of chemicals having identical Mold descriptor sets (point color differentiates chemical).

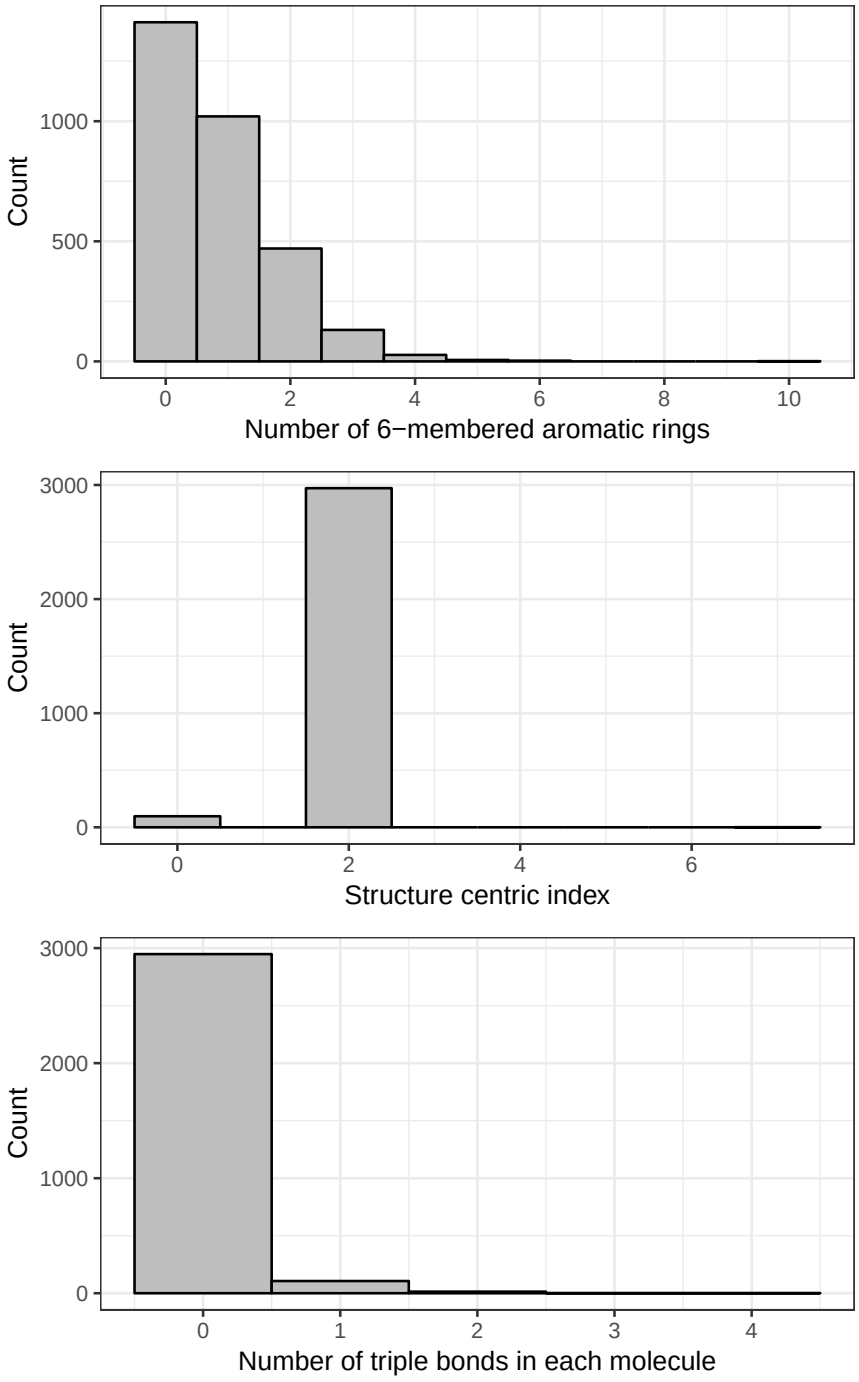

Fig S.21: Three examples of count data in the set of Mold2 descriptors X for the included chemicals. Note that some features seem better suited to the underlying continuous assumption used in the model (i.e., the top and bottom) while others could be improved on via some other special specification based on expert input (i.e., the middle).

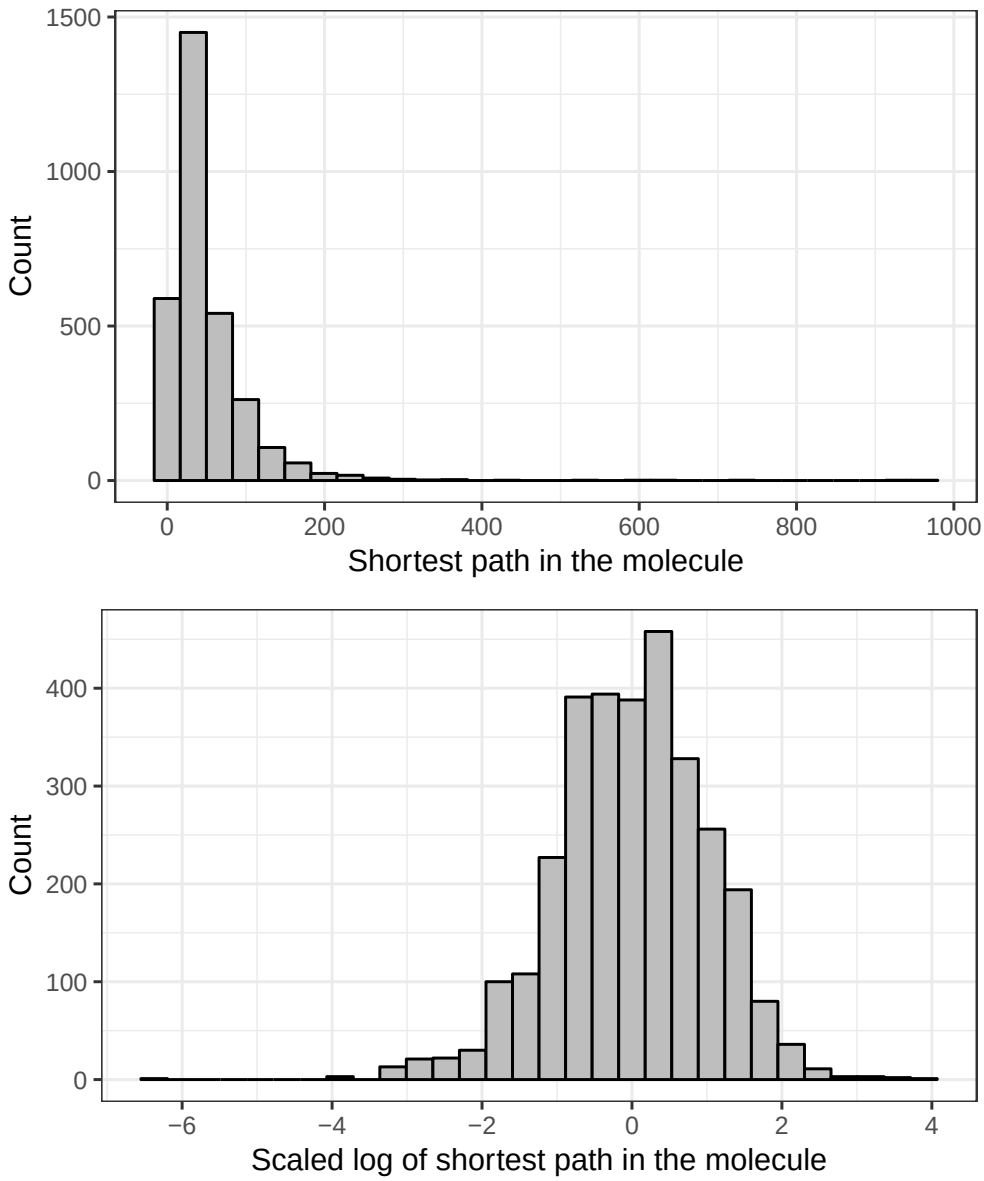

Fig S.22: Count variables having maximum value of greater than 10 are log transformed and then treated as continuous (i.e., scaled and centered) before their inclusion in the model. The top row shows an example of a pre-transformed feature, and the bottom row shows that same feature after taking the log, scaling, and centering.