Published in final edited form as:
*Environ Sci Technol*. 2019 January 15; 53(2): 719–732. doi:10.1021/acs.est.8b04056.

# Consensus Modeling of Median Chemical Intake for the U.S. Population Based on Predictions of Exposure Pathways

**Caroline L. Ring**[1,3,§], **Jon A. Arnot**[4,5,6], **Deborah H. Bennett**[7], **Peter P. Egeghy**[2], **Peter Fantke**[8], **Lei Huang**[9], **Kristin K. Isaacs**[2], **Olivier Jolliet**[9], **Katherine A. Phillips**[2], **Paul S. Price**[2], **Hyeong-Moo Shin**[10], **John N. Westgate**[4,°], **R. Woodrow Setzer**[1], **John F. Wambaugh**[1,*]

[1]National Center for Computational Toxicology, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711

[2]National Exposure Research Laboratory, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711

[3]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee 37831

[4]ARC Arnot Research and Consulting, 36 Sproat Ave. Toronto, ON, Canada, M4M 1W4

[5]Department of Physical & Environmental Sciences, University of Toronto Scarborough 1265 Military Trail, Toronto, ON, Canada, M1C 1A4

[6]Department of Pharmacology and Toxicology, University of Toronto, 1 King's College Cir, Toronto, ON, Canada, M5S 1A8

[7]Department of Public Health Sciences, University of California, Davis, California, 95616

[8]Quantitative Sustainability Assessment Division, Department of Management Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[9]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109

[10]Department of Earth and Environmental Sciences, University of Texas, Arlington, Texas, 76019

## Abstract

Prioritizing the potential risk posed to human health by chemicals requires tools that can estimate exposure from limited information. In this study, chemical structure and physicochemical properties were used to predict the probability that a chemical might be associated with any of four exposure pathways leading from sources–consumer (near-field), dietary, far-field industrial, and far-field pesticide–to the general population. The balanced accuracies of these source-based

[*]Corresponding Author: John F. Wambaugh, 109 T.W. Alexander Dr, NC 27711, USA, Wambaugh.John@epa.gov, Phone: (919) 541-7641.
[§]Current Address: ToxStrategies, 9390 Research Blvd #100, Austin, TX 78759, Dr. Ring received no funding from ToxStrategies or any of its clients for this project, and neither ToxStrategies nor any of its clients was involved in the development or approval of this research or this report.
[°]Current Address: Pest Management Regulatory Agency, Health Canada, Ottawa, ON, Canada
Author Contributions
The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript

exposure pathway models range from 73 to 81%, with the error rate for identifying positive chemicals ranging from 17 to 36%. We then used exposure pathways to organize predictions from 13 different exposure models as well as other predictors of human intake rates. We created a consensus, meta-model using the Systematic Empirical Evaluation of Models framework in which the predictors of exposure were combined by pathway and weighted according to predictive ability for chemical intake rates inferred from human biomonitoring data for 114 chemicals. The consensus model yields an R2 of ~0.8. We extrapolate to predict relevant pathway(s), median intake rate, and credible interval for 479 926 chemicals, mostly with minimal exposure information. This approach identifies 1880 chemicals for which the median population intake rates may exceed 0.1 mg/kg bodyweight/day, while there is 95% confidence that the median intake rate is below 1 μg/kg BW/day for 474572 compounds.

## Introduction

One measure of the risk to human health posed by chemicals in the environment is the ratio of the dose that potentially causes adverse health effects (i.e., hazard) to the dose received from the environment (i.e., exposure)[1]. Unfortunately, existing sources of hazard and exposure data do not address many thousands of chemicals that may be present in the environment or used in commerce[2–6]. High-throughput methods allow the identification and further testing of those chemicals which are more likely to pose a risk to people. Comparing hazard and exposure is most meaningful when the uncertainty in these estimates is quantified, thereby allowing comparison between the upper bound on exposure and lower bound on hazard. This approach is being considered by the governments of Australia, Canada, Europe, Japan, Korea, Singapore, and the United States as a tool to help accelerate the “pace of chemical risk assessment”[7]. Despite relatively large uncertainties, the estimated ratio of hazard to exposure exceeds one million for many chemicals, which allows separation of those chemicals for which risk is plausible from those chemicals that are less likely to pose a potential health risk[8]. We distinguish between probabilistic risk-based chemical priority setting and formal risk assessment; the latter is more resource-intensive but may substantially revise the estimated margin between hazard and exposure[1, 3, 9].

The EPA’s Exposure Forecasting (“ExpoCast”) project relies upon high-throughput exposure (HTE) models requiring minimal chemical-specific information[10] to provide risk-based context for the high-throughput *in vitro* bioactivity data that are now available for large numbers of chemicals[11–13]. HTE models are used to make quantitative predictions of chemical intake based upon the mechanism of exposure and description of the exposure event. Each HTE model can describe one or more exposure pathways, which the International Society for Exposure Science defines as “The course an agent takes from the source to the target.” [14] Model pathways involve a chemical source (e.g., a household product), interactions with environmental media (e.g., air, water, indoor surfaces), and a target (e.g., a human individual)[15]. Each HTE model may also reflect differing assumptions, data for describing the environment, choices of evaluation data, and a range of criteria for “acceptable” performance[16]. In addition to these mechanistic HTE models, ExpoCast also uses other relevant chemical-specific metrics that may be predictive of exposure (for example, existing regulatory exposure assessments or production volumes). Here, both HTE model predictions and other chemical-specific metrics (collectively referred to here as “predictors”) are systematically evaluated to assess the relative and collective contributions to predict exposure.

To integrate, evaluate, and calibrate existing exposure predictors, ExpoCast has developed an approach called the Systematic Empirical Evaluation of Models (SEEM) framework. With SEEM, the available predictors are combined into a consensus[17] Bayesian regression meta-model for median population exposure via comparison with human biomonitoring data. SEEM is a “meet-in-the-middle” approach[18] in which predictions of chemical intake rates from “forward” HTE models are compared with rates inferred from the Centers for Disease Control and Prevention’s (CDC) National Health and Nutrition Examination Survey (NHANES) biomarker data using “reverse” models that attempt to reconstruct exposures. In the resulting meta-model, predictors are weighted to reflect their ability to predict the biomonitoring-based intakes, and the residual error in the regression provides an estimate of the model uncertainty. Here, we improve upon the existing SEEM regression models by incorporating exposure pathway information into the model structure.

Incorporating pathway information into SEEM is critical, as not all predictors are relevant for all pathways and chemicals. Here we use the simple term “pathway” to represent the totality of paths that a chemical may follow from a particular *source-type* to reach a person. Individual source-types, or pathways, contribute widely-varying amounts to overall exposure. Proximate sources of exposure (i.e., in ‘near-field’ consumer use scenarios) are often the largest contributors to exposures for chemicals with biomonitoring data collected by NHANES, while distant sources of exposure generally contribute less to overall exposure (i.e., in ‘far-field’ scenarios wherein individuals are exposed to chemicals that were released or used away from the current exposure event) [19–21]. In updating the SEEM meta-model approach, we consider four source-based pathways — consumer (near-field), far-field pesticidal, non-pesticidal dietary, and far-field industrial.

Here we have developed an updated consensus SEEM meta-model by applying the revised, pathway-based SEEM approach to a suite of thirteen HTE predictors from ExpoCast and its collaborators. Each of these predictors is relevant to one or more of the four exposure pathways included in the model. The new resulting consensus meta-model is then applied to predict chemical intakes for a library of 681,609 chemicals. Since exposure sources and pathways are unknown for the majority of these chemicals, we also build chemical structure- and property-based machine learning models that quantify the relevance of a chemical to the four pathways studied here. Integrating the results of these “pathway predictor” models with the SEEM meta-model provides the first estimates of potential human exposure the majority of chemicals analyzed.[6] These estimates and their uncertainty allow human health risk-based comparisons with toxicity data for chemicals subject to the four pathways described [8, 22].

## Materials and Methods

The SEEM approach evaluates predictors of exposure based on how well they correlate with estimates of chemical intake rate from biomonitoring. Evaluation chemicals and predictors are now grouped according to relevant exposure pathway(s). The predictors are combined into a consensus model, with each predictor weighted according to its empirically estimated predictive ability. Any errors, approximations, and/or assumptions act to increase the estimated uncertainty. A technical glossary of terminology is provided in Table 1.

As illustrated in Fig 1, the updated SEEM framework uses linear regression (Equation 1) to integrate all the exposure predictors in Table 2 into a consensus meta-model. When given the pathway predictions and other inputs this meta-model can then provide to provide a consensus prediction. This method is analogous to quantitative structure-property predictors[23]. Each step is described in detail below, but in summary: Bayesian methods are used to estimate possible parameters for Equation 1 that would make the exposure predictors consistent with intake rates inferred from NHANES biomonitoring data for 114 chemicals. Both the exposure predictors and the NHANES chemicals are assigned to one or more pathways. For chemicals without NHANES data, machine learning methods are trained and used to assign chemicals to pathways based upon structure, after which Equation 1 is used to make predictions for a large library of chemical structures.

All analyses were performed in R version 3.4.3 [24]. All chemical descriptors, predictors, pathway information, and meta-model predictions are provided in Supplemental Table 1. All chemical predictions are available through the EPA Chemistry Dashboard (https:/comptox.epa.gov)[25].

## Materials and Methods (MM) Section 1: Linear Regression Model

The intake rates ($R_i$) in units of mg/kg bodyweight (BW)/day for chemical *i* were approximated as:

$$\boldsymbol{\log R_i = a_0 + \sum_j \delta_{i,j} \times \left(a_j + \sum_k w_{j,k} \times \log \pi_{i,k}\right)} \quad \text{Equation 1}$$

where $a_0$ represents a “grand mean” intake rate over all chemicals that is unexplained by the exposure predictors, $\delta_{ij}$ is a Boolean (0/1) value that represents whether or not exposure to chemical *i* occurs via pathway *j*, $a_j$ represents the additional mean intake rate via pathway *j* over all chemicals with exposure via pathway *j* that is unexplained by the predictors, and $w_{j,k}$ represents the loading (“weight”) of predictor k for chemical *i* by pathway *j* ($\boldsymbol{\pi}_{i,k}$). $w_{j,k}$ is zero if an exposure predictor does not correspond to pathway *j*. The exposure predictors and intake rates inferred from NHANES data were scaled as $X' = \frac{X - \overline{X}}{sd(X)}$ so that their value indicates the number of standard deviations ($sd(X)$) above or below the average ($\overline{X}$) for predictor or rates *X*. This scaling makes the $w_{j,k}$ directly comparable to each other.

## MM2: Bayesian Inference of Linear Model Parameters

The SEEM approach evaluates predictors of exposure based on how well they correlate with estimates of chemical intake rate. The regression model means $a_j$ and loadings $w_{j,k}$ were estimated from inferred NHANES intake rates using Bayesian analysis. Bayesian analysis combines observed data with prior information and assumptions about the process that created the data (i.e., a statistical model) to estimate correlated distributions of model parameter values that are consistent with the observed data, prior information, and structure of the statistical model [26]. Here, the parameter distributions for the predictor loadings reflect different possible combinations of models and other exposure predictors. For example, perhaps model A is very important in predicting exposure and model B provides almost no information; or perhaps models A and B are equally informative. By sampling from the distributions of parameters and evaluating the model with each sampled parameter set, a distribution of predictions can be made which can be characterized by a 95% "credible interval". Predictions have 95% probability of falling within the 95% credible interval. (The Bayesian concept of a "credible interval" is analogous to, but distinct from, the frequentist concept of a "confidence interval".[26])

Bayesian analysis was performed using Markov Chain Monte Carlo (MCMC) as implemented by JAGS [27] version 4.2.0, accessed through the R package "runjags" [28]. Statistical model code is provided in the Supplemental Material. Observations (the inferred population median NHANES intake rates) were assumed to be log-normally distributed about the SEEM meta-model prediction $R_i$ (Equation 1), with a standard deviation that included a chemical-specific (observation) error $\sigma_i$ (reflecting the uncertainty in the population median) and a SEEM meta-model error $\sigma_o$, so the total error for a chemical was $\sigma_{total,\, i} = \left(\sigma_0^2 + \sigma_i^2\right)^{1/2}$. The grand and pathway-specific means ($a_0$ and $a_j$) were assumed to obey a multivariate normal distribution. The nonzero loadings $w_{j,k}$ on the centered and scaled model predictions were also sampled from a multivariate normal distribution. For both the means and loadings, the mean of the multivariate distributions was assumed to obey a multivariate normal hyper-prior, while the correlation matrices were assumed to obey Wishart distributions. The performance of the calibrated model is approximated using the base R function "lm" (linear model) to calculate $R^2$ and root mean squared error (RMSE) values for a linear regression on the median inferred exposures.

## MM3: Pathway Prediction Models

The NHANES chemicals were manually curated using CPDat and assigned definitively as a "positive" or "negative" reference chemical for each pathway (Supplemental Table 2). To quantify the probability that each pathway is relevant to other chemicals, the random forest algorithm [29, 30] was used for each pathway in turn. The random forest algorithm is a probabilistic extension of a standard decision (classification) tree approach [31]. For chemicals with data about whether or not exposure occurs via a specified pathway, a standard classification tree approach would find a set of rules for dividing these chemicals into groups based on combinations of their physical-chemical properties and structural descriptors to produce groups that are as homogeneous as possible with regard to whether exposure occurs via the specified pathway. The most-common answer in each group (yes, exposure occurs via this pathway; or no, exposure does not occur via this pathway) is considered the tree-predicted answer for all chemicals in that group. Then, the tree rules can be applied to chemicals without pathway data, to predict whether or not the specified pathway is relevant for them. The random forest algorithm extends the classification tree approach by using a large ensemble of classification trees, each trained on a randomly-chosen subset of the available data. For each pathway, 5,000 trees were used. Each tree produced its own model-predicted yes/no classification for whether exposure occurred via the specified pathway for each chemical. The fraction of trees for pathway $j$ that predict a "yes" for chemical $i$ can be interpreted as the probability that exposure to chemical *i* occurs via pathway *j*. The performance of each classifier tree can then be estimated using the data withheld from the training set for that tree, the so called "out of bag" (OOB) error [29, 30]. Models were evaluated using mean OOB error.

In a standard classification tree, chemicals are repeatedly split into more-refined groups in an iterative fashion. Starting with all chemicals in one group, the tree algorithm considers splitting the group based on each of the predictors in turn and chooses the split that will result in the most homogeneous possible groups. The process repeats iteratively for each resulting group until the tree reaches some pre-determined metric balancing accuracy and complexity. The random forest randomizes the decision-tree algorithm in that each tree in a random forest considers not all possible predictors, but only a randomly-selected subset of the available predictors each time it chooses the best possible split [32]. The importance of a predictor is then assessed based on the increase in group homogeneity resulting from splitting on that predictor, averaged across all trees in the random forest [33].

To train the random forest algorithm for each pathway, we chose sets of chemicals that might reasonably serve as "positive" or "negative" examples of that pathway, as summarized in Table 3. The number of positive and negative chemicals was balanced by randomly sampling a subset from the larger of the two sets. Chemicals on the FDA Cumulative Estimated Daily Intake (CEDI) list were considered positives for the dietary pathway. The ExpoCast screen of chemicals in household products and the Chemical and Products Database (CPDat) provided information for multiple sources. The U.S. Geological Services National Water-Quality Assessment data portal was used for compounds observed in water to identify chemicals with far-field industrial sources. Finally, the EPA Chemistry Dashboard (http://comptox.epa.gov) was used to obtain lists of per- and poly-fluorinated chemicals from the NORMAN Network (the Network of Reference Laboratories, Research Centres and Related Organisations for Monitoring of Emerging Environmental substances).

The random forest algorithm used a total of 743 chemical descriptors as potential predictors of exposure pathways. Chemical structures were used to predict 13 physico-chemical properties using the OPEn structure–activity/property Relationship App (OPERA) models [34] for water solubility, vapor pressure, hydroxylation rate constant for the atmospheric, gas-phase reaction, bioconcentration factor, biodegradation half-life, anaerobic biodegradability, boiling point, Henry's Law constant at 25 °C, fish whole body biotransformation rate constant, octanol:air partition coefficient, octanol:water partition coefficient, octanol:soil organic carbon partition coefficient, and melting point (10/27/2016)[34]. For convenience in working with large numbers of chemicals and to evaluate the uncertainty introduced when predictions are used for chemicals without measured data, only predicted values were used. Predicted values might increase the overall estimated error rate (Table 3). Molecular weight was considered as an additional descriptor, as calculated directly from formula by EPA's Distributed Structure-Searchable Toxicity (DSSTox) Database [35]. Finally, 729 binary ToxPrint descriptors were used to identify the presence (1) or absence (0) of specific chemical substructures within each structure [36]. ToxPrint descriptors are open-source (freely available) descriptors that are "designed to provide excellent coverage of environmental, regulatory, and commercial-use chemical space"[36], as opposed to proprietary and/or pharmaceutical-centered chemical descriptors [37].

## MM4: Meta-Model Predictions and Uncertainty Quantification

Three items of information are needed to make SEEM exposure forecasts (Figure 1): 1) Markov chains describing likely sets of linear regression model parameters (MM2), 2) estimates of whether a given pathway *j* is relevant to a given chemical *i* (Probability($\delta_{ij}$), MM3), and 3) any available predictors (MM6). Because the linear regression is conducted on scaled and centered predictors, chemicals without a prediction for a specific model are assigned the average value for that model. This average value only contributes to the overall predicted rate if the $\delta_{ij}$ for the chemical and pathway is 1. When chemicals are not assigned to any pathway, their predicted exposure rate is $a_0$. Predictions were made for a total of 687,359 compounds for which the chemical structures could be obtained from DSStox. Example calculations using the medians of the regression model parameters distributions are given for the NHANES chemicals in Supplemental Table 4. In practice, the inferred exposure rates are used rather than the meta-model predictions for the NHANES chemicals.

Errors, approximations, and assumptions all act to increase the residual differences between the predicted and inferred intake rates. Uncertainty is characterized via the Bayesian analysis: the parameters for Equation 1 that are more uncertain have larger estimated distributions. A chemical-specific credible interval for each prediction is calculated using 500 sets of coefficients for Equation 1 that are sampled from the Markov chain (Section MM2). For each draw the values $\delta_{ij}$ are assigned from a Bernoulli distribution with the probability predicted in Section MM3. The median and quantiles for 0.025 and 0.975 (the 95% credible interval) are calculated from the 500 draws. The mean and standard deviation of the NHANES observations are used to transform the scaled predictions back to mg/kg BW/day.

## MM5: Chemical Intake Rate Inferences

Median population chemical intake rates were inferred from biomonitoring data. Data on urine were previously analyzed and published [20, 22], while data on serum and blood concentrations are newly analyzed here.

Biomonitoring data were collected by NHANES, a rolling survey covering roughly 10,000 individuals every two years. NHANES uses a deliberate sampling protocol so individuals can be weighted to reflect the U.S. population. Urine concentration data were analyzed by Wambaugh et al. [20] using a model that assumed the concentration of an analyte in urine represented steady-state exposure(s) to one or more parent compounds. As chemicals are sometimes removed from NHANES monitoring, the most recently reported two-year period between 1999 and 2010 was used for each chemical. These estimated exposures were reported in Supplemental Information by Ring et al. [22].

Serum and blood concentrations (*Conc*) were related to intake rates assuming steady-state exposure. Chemical-specific whole body clearance rates (CL, in units of L/kg BW/hour) were estimated using the R package "httk" v1.8 [38], which included *in vitro* measured toxicokinetic rates [8, 39, 40]. Exposures were estimated as $Rate = Conc \times CL$. A small subset of NHANES chemicals have human half-lives reported in Arnot, et al. [41]. The clearance rates for two perfluorinated chemicals (PFOS and PFOA) measured in NHANES were taken from recent chemical-specific assessments (0.000081 and 0.00014 L/kg BW/day).[42, 43]

## MM6: Exposure Predictors (High-Throughput Exposure Model Predictions and Metrics)

The consensus intake rate model (Equation 1) is a meta-model that weighs various exposure predictors according to their ability to predict the available data on intake rates. Exposure predictors are organized via the pathways shown in Figure 2. Each predictor is described briefly below, with additional information provided in the associated references. Chemical descriptors were harmonized for the various HTE models by providing the same chemical descriptor set for analysis. Since there are limitations on the number of chemicals that can be run by some models, only those chemicals predicted with >80% chance of being relevant to a pathway were provided. The pathway predictor models were refined somewhat while the HTE model predictions were being generated. Chemicals were described using the OPERA [34] physico-chemical descriptors (10/27/2016).

**MM6.1: Production Volume—**Information on chemical production volume was obtained from the 2015 EPA Chemical Data Reporting (CDR) under the Toxic Substances Control Act (TSCA) [44]. Given that production volumes are provided in coarse bands (e.g., 1M-10M lb./year), the geometric mean of the limits of each bin was used. Further, the production volume of many compounds was confidential, in these cases the average production volume of chemicals with reported CDR production volume was used. While 7856 compounds were covered by the CDR, compounds which were not on the list were assumed at 12,500 lb/year, half the minimum requirement for being listed. Chemical production volume was considered as a potential predictor for exposure for all four pathways.

**MM6.2: Stockholm Convention—**The Stockholm Convention on Persistent Organic Pollutants lists persistent and/or bioaccumulative pollutants whose production is being banned or reduced by international treaty [45]. The list includes specific chemicals as well as broad categories (e.g., all polychlorinated biphenyls or PCBs). Manual enumeration was used to identify all chemicals within each class. This list is provided as Supplemental Table 3. Presence on the Stockholm Convention list was examined as a predictor for industrial and pesticidal pathways.

**MM6.3: Intake Rate Estimates from EPA REDs—**As a part of ongoing assessment of pesticide safety, the EPA generates re-registration eligibility documents (REDs) that include estimated intake rates for the general U.S. population (mean). These assessments have been collected by Wetmore et al. [8] as of 2015 and provide predictions for chemicals with far-field pesticide sources.

**MM6.4: Food Contact Migration Exposure Model—**The model of Biryol, et al. [46] makes predictions of migration of chemicals into food from packaging (and resulting food concentrations) based upon physico-chemical properties, packaging formulation, and the properties of the food. The model was developed by fitting a parsimonious linear model to measured migration data obtained from the FDA [47]. These migration estimates were combined with food intakes from NHANES to estimate exposures via the dietary pathway (mg/kg BW/day) for the U.S. population for 1009 chemicals publicly listed by U.S. or European regulatory agencies as potential polymer food contact materials. These predictions, available in the supplemental information of Biryol, Nicolas, Wambaugh, Phillips and Isaacs [46], were used for the dietary pathway.

**MM6.5: Fate and Transport Models—**Fate and transport models predict intake rates based upon compound-specific data both on distribution (e.g., using fugacity) and degradation (i.e., using half-lives in environmental media). In many cases, these models solve the underlying mass balance at steady-state. Such models usually account for multimedia fate and multi-pathway exposure and range from generic nested models like USEtox[48] to spatially explicit multiscale models like Pangea [49]. These models predict chemical fate and distribution in representative environmental media (e.g., air, water, soil, sediment, biota). “Exposure factors” describing human interaction with environmental media and diet are used to determine an average predicted intake rate per each kg of chemical emitted. [48, 49].

Since the models for far-field sources make predictions for intake rates based upon the amount emitted (i.e., mg/kg BW/day per kg emitted), knowledge of rate and media of release to the environment is needed. Unfortunately, this information is not available on a high-throughput scale. Instead, production volumes (MM6.1) were used as an additional predictor of exposure ($\pi_{i,k}$ in Equation 1). Because the models are evaluated on the log scale (that is, for two model predictions pred1 and pred2, log rate = log pred1 + log pred2) and the sum of two logarithms is equal to the product, (i.e., log pred1 + log pred2 = log[pred1 * pred2]), we evaluate the effectiveness of production volume as a surrogate for environmental release.

**MM6.5.1: USEtox Model:** The United Nations Environment Programme (UNEP) and Society for Environmental Toxicology and Chemistry (SETAC) toxicity and ecotoxicity characterization model USEtox [48] version 2.0 is a global scientific consensus fate, exposure and effect model that was used as a predictor of far-field industrial sources of chemical exposure. USEtox 2.0 consists of a set of nested environmental compartments at indoor, urban, continental, and global scale.

**MM6.5.2: dynamiCROP pesticide exposure model parameterization:** The dynamic plant update and crop residue exposure model dynamiCROP [50, 51] was implemented in USEtox version 2.0 in a parameterized version [52, 53], and used as a predictor of far-field sources of pesticide exposure. The model starts from a set of pesticide mass fractions reaching crop, soil, and air upon pesticide application and predicts pesticide residues in crops at any given harvest time. Human exposure is then linked to the consumption of these residues in harvested crop components after undergoing food processing, and inhalation and ingestion exposure is associated with the fractions lost to air and soil after far-field environmental distribution.

**MM6.5.3: RAIDAR Model:** The Risk Assessment IDentification And Ranking (RAIDAR) model is an environmental fate and transport model linked with food web bioaccumulation models for representative ecological and agricultural targets and humans. RAIDAR was used for far-field industrial chemicals and pesticides. [54]

### MM6.6: Consumer (Near-Field Source) Models—

All models covering the consumer pathway require estimated release rates for individual chemicals from consumer products. CPDat [55] and data on consumer product use in the high-throughput Stochastic Human Exposure and Dose Simulation model (SHEDS-HT) [56] were used to generate inputs (product-specific releases) representative of the median U.S. population. These releases were aligned with appropriate near-field compartments (e.g., air, surfaces, skin) in each model and summed across products to estimate the total release for each chemical. The SHEDS-HT model [56] includes a parameterization of the indoor environment that combines data on which chemicals are in what products (i.e., composition data) with how often these products are used (i.e., pattern of consumer product use) to generate indoor chemical releases (g/day).

**MM6.6.1: SHEDS-HT:** In SHEDS-HT, the residential module of SHEDS-Multimedia [57] was modified to reduce the user burden, input data demands, and run times of the higher-tier model. SHEDS-HT links chemicals to consumer product categories or food groups (and thus exposure scenarios) to make predictions for both direct (intentional use) and indirect exposure for the consumer pathway. In modeling indirect (post-use) exposures from near-field sources, SHEDS-HT employs a fugacity-based module to estimate concentrations in indoor environmental media. For direct exposures, SHEDS-HT calculates route-specific (dermal, inhalation, and ingestion) and total exposure from each product type modeled, and then calculates an aggregate chemical exposure from all products for each chemical. In this study, SHEDS-HT used the latest version of CPDat [55] and release v0.1.6 of the SHEDS-HT code (https://github.com/HumanExposure/SHEDSHTRPackage/releases/tag/v0.1.6) and default input files.

**MM6.6.2: FINE Model:** The Fugacity-based INdoor Exposure (FINE) model is a near-field fate and exposure model for organic compounds released to the indoor consumer environment [58, 59]. The model simulates the concentrations of organic compounds in various indoor compartments (e.g., gas phase, airborne particles, dust, carpet, vinyl flooring, and walls). The model estimates the intake fraction due to indoor air releases through inhalation, dermal uptake, and non-dietary dust ingestion by coupling the simulated concentrations with recommended exposure factors (e.g., inhalation rate, dust ingestion rate), which were obtained from the EPA Exposure Factors Handbook[60].

**MM6.6.3: RAIDAR-ICE:** The RAIDAR-Indoor and Consumer Exposure (RAIDAR-ICE) model [61] combines an indoor fate model with a physiologically-based biokinetic model for simulating exposures and potential risks to a representative human adult living in the indoor environment. RAIDAR-ICE is based on the Indoor Chemical Exposure Classification/Ranking Model (ICECRM) [62], but in addition to the indirect inhalation, non-dietary ingestion, and dermal exposure pathways included in ICECRM, RAIDAR-ICE also includes direct inhalation, ingestion, and dermal exposure pathways.

**MM6.6.4: Product intake fraction modeling framework:** For considering consumer exposure in USEtox, the product intake fraction is used as a metric linking the mass taken in by humans via all pathways per chemical mass in a specific product application[63]. Potential input models to estimate product application-specific pathways and processes are summarized in Huang et al. [64], with specific models coupled with USEtox already for building materials [65], food contact materials [66] and personal care products [67].

These exposure predictors correspond to the USEtox Consumer Scenario and USEtox Dietary Scenario in Table 2.

## Results and Discussion

### Source-based Pathway Predictor Models

The 2017 report by the U.S. National Academy of Sciences, Engineering, and Medicine (NASEM) "Using 21st Century Science to Improve Risk-Related Evaluations" identified the need for HTE models as the basis for "exposure-based priority setting". These models enable a high throughput chemical risk estimation approach under evaluation by regulatory agencies worldwide [7, 68], in which dose rates predicted to cause *in vitro* bioactivity are compared directly to predictions of intake rate [1, 8]. However, Shin et al.[69] argued that, in the absence of chemical-specific knowledge of relevant pathways, chemical exposures should be simulated via all pathways, generating overly conservative (i.e*.,* higher than reality) estimates of exposure [70]. The research reported here advances the exposure models available for quantitative risk estimation by predicting the pathways likely relevant to each chemical based upon structural features and physico-chemical properties and then using this information to incorporate a wider range of exposure predictors (Table 2).

Machine learning models were built for each of four source-based pathways – far-field pesticide use, non-pesticide dietary exposure, far-field industrial exposure (for example, via drinking water), and consumer (for example, near-field exposure to household products). The performance of the random forest models for making chemical-specific exposure pathway predictions are summarized in Table 3. The pathway model OOB error rates range from 19–27%. The balanced accuracies range from 73–81%, with the error rate for identifying positive chemicals ranging from 16–36%. Generally, the physico-chemical properties and molecular weight were predictive to varying degrees for all pathways. The ToxPrint structure descriptors were more mixed in their importance for predicting pathways, which is expected since any one structural feature is only present in a small subset of chemicals (see Supplemental Table 2).

The pathway predictors are developed by determining training sets of reference chemicals that are either known to have exposure via a given pathway ("positives") or known to not have exposure via that pathway ("negative"). Techniques such as suspect screening analysis (SSA) [71] and non-targeted analysis (NTA)[72] offer the promise of identifying more reference chemicals for pathway predictive models, as in Phillips, et al.[73], whose SSA measurements were used here to identify chemicals with consumer pathway exposure. These models can then be informed via SSA/NTA of relevant media, such as water, in which the chemicals that were checked for (via analytical standards) but found missing become negatives assuming that the limits of detection for that medium are acceptable [71, 74]. However, negatives are still problematic since it is possible that a chemical not found by gas chromatography might be found by liquid chromatography, and *vice versa*. One route to developing better training sets would be to subdivide the pathways so that each contained more homogenous chemicals where one can be confident that the SSA/NTA method employed should detect the chemicals, if present. Pharmaceuticals provided helpful negatives for these four pathways, but do have relevant exposure both from usage as pharmaceuticals, and environmental release [75]. If monitoring data similar to NHANES on a significant number of pharmaceuticals becomes available, it would allow for the addition of pharmaceutical-relevant pathways. [55, 71, 74, 76].

## Development of the Pathway-Based SEEM Meta-Model

We use the SEEM approach to construct a meta-model of predictors based on their ability to describe exposures inferred from the NHANES data. Pathway-based analysis allows organization of predictors and evaluation chemicals based on exposure pathways. Models that describe, for example, outdoor environmental fate and transport are only examined for predictive ability against chemicals with some exposure contribution from outdoor environments. In addition to models, data on chemical production volume are analyzed as exposure predictors, as well as an indicator (0 or 1) of whether chemicals are listed as persistent organic pollutants by the Stockholm convention. Finally, actual intake rate estimates from pesticide REDs are included as exposure predictors – these predictions are evaluated against inferred intake rates in the same way as the other predictors and are included in the consensus model predictions.

Both serum and urine data were used to infer 114 chemical intake rates from NHANES (plotted in Supplemental Figure 1 and described in Supplemental Table 4), with mean 0.16 ng/kg BW/day and standard deviation of 132x greater or lesser. These rates spanned roughly twelve orders of magnitude (from $10^{-15}$ mg/kg BW/day for 2,2’,4,4’,5,5’-hexachlorobiphenyl to 0.001 mg/kg BW/day for diethyl phthalate) representing a significant increase from the six orders of magnitude range inferred from urine alone by Wambaugh et al. (2014) [20]. Biomonitoring reflects integration of all routes of exposure and some NHANES chemicals are involved in multiple pathways. NHANES covers many pesticides and chemicals with consumer exposures but has lesser coverage of chemicals with industrial (far-field) and dietary sources. The distribution of NHANES chemicals among the pathways is listed in Table 3.

The exposure pathway indicators ($\delta_{ij}$) were used to allow multivariate regression of inferred intake rates on the appropriate predictors. The pathway means indicate relative changes in intake rate associated with chemicals for each pathway (Figure 2): +1.02 (consumer), +0.707 (dietary), +0.572 (far-field pesticides) and −0.08 (far-field industrial). Only the consumer and dietary pathway are significantly non-zero, with fold-changes of 145 times greater intake rates for consumer pathway exposures and 21 times higher for dietary exposures relative to the overall mean $a_0$ of 1.2 ng/kg/day. Chemicals with proximate sources were higher than chemicals with only far-field sources, recapitulating the finding from the previous SEEM “heuristic” model [20]. In contrast, chemicals with only far-field sources are lower than average. The indicator variable for presence on the Stockholm Convention list of persistent chemicals was a statistically significant negative factor for both pesticides and industrial chemicals, with banned pesticides being 1670 times lower, and banned industrial compounds being 5134 times lower on average. Median estimates for all coefficients in the multivariate regression are given in Supplemental Table 6. The SHEDS-HT Direct, Food Contact, USEtox (pesticide only), and production volume (pesticide only) predictors showed significant positive correlation with increased intake rate. The RAIDAR (industrial only), USEtox (diet only) and FDA CEDI predictors all showed significant negative correlation with intake rate.

Because correlations exist between the predictors it is difficult to directly interpret the contributions of each model to the consensus prediction. A univariate (one predictor at a time) analysis was performed to determine the correlation between each predictor and chemicals associated with a given pathway (Supplemental Figure 2). The results for all predictors are given in Supplemental Table 7. In the univariate analysis, the SHEDS-HT Direct pathway, RAIDAR-ICE, and production volume all correlate positively (i.e., 95% probability of a non-zero dependence) with intake of chemicals by the consumer pathway. Both the USEtox and Stockholm Convention predictors correlate negatively with intake rate from the far-field pesticidal sources, while production volume correlates positively. Only the Stockholm Convention correlates (negatively) with far-field pesticide exposure. Only the Food Contact model correlates (positively) with dietary exposure. All other predictors were not statistically significant in the univariate analysis; however, while the median estimated loading coefficients $w_{ik}$ for the meta-model are near zero, the actual values may be positive or negative, given the limitations of the evaluation chemicals. Notably, some predictors that are significant in the multivariate model do not demonstrate significance in the univariate analysis, indicating the benefit for including many relevant models in the meta-analysis[17].

In Figure 3, we examine the correlation between inferred intake rate from the NHANES biomonitoring data and SEEM meta-model prediction. First, we note that the $R^2$ ~0.8 for a weighted linear regression using median coefficient values (trend-line Figure 3) is an improvement over the predictive ability of the Wambaugh et al. [20] empirical exposure model ($R^2$ ~0.5). The RMSE for the model is ~0.93.

We note that all but six chemicals in Figure 3 have inferred daily intake rates greater than $10^{-9}$ mg/kg BW/day. These six chemicals consist of five pesticides (Endrin, Mirex, Aldrin, Dieldrin, and p,p'-DDT) and PCB-153. All six chemicals are persistent organic pollutants listed by the Stockholm convention that are detectable in serum by NHANES. The inferred intake rates for these six chemicals range from $9\times10^{-16}$ to $2\times10^{-13}$ mg/kg BW/day. When these six chemicals are omitted, the $R^2$ and RMSE for the linear regression in Figure 3 changes only slightly to 0.82 and 0.9, respectively. However, the $R^2$ for the six chemicals is only 0.03 (RMSE 13.5), indicating that better modeling of these persistent chemicals that have very low intake rates is needed. Uncertainty in "emission rates" assumed for these persistent chemicals here is a likely explanation for the poor agreement since human exposure model evaluations with more detailed emissions information generally show much better agreement, e.g. Li et al.[77]. The present consensus model predictions are biased conservatively, in that they tend to overestimate intake rates for these six chemicals. (The dashed line in Figure 3 indicates the 1:1 (perfect predictor) – any point over the 1:1 line is overpredicted.)

## Consensus Model Predictions

For 687,359 chemicals with structures curated by the EPA's DSStox library, we used the pathway predictor models to identify relevant pathway(s) and the consensus meta-model to predict the population median intake rate with 95% credible interval. In Figure 4, exposures are predicted for 479,926 chemicals. The pathway predictor models are independent and uncorrelated, so that a chemical could have exposure via multiple pathways or none. Each of the thousands of pathway training set chemicals (Table 3) was definitively assigned either 100% or 0% probability depending on whether they were positive or negative training chemicals. Note that the x-axis on the left-hand side of Figure 4 is logarithmic. On the left-hand side of Figure 4, each chemical's plot symbol is assigned based on whether there is an 50% or greater probability of exposure by each pathway. The credible interval reflects a combination of pathway assignments, pathway means, and model predictions and loadings. See Supplemental Table 1 for all chemical descriptors and predictions. Given that most chemicals do not have model predictions (Table 2), chemicals intake rates are only being driven by the pathway means based on the probability of those pathways being relevant to the chemical.

With source-based pathway annotation, we have gained additional information about the likely pathway(s) of exposure, and the ability to augment average pathway exposures with calibrated model predictions. Of the chemicals examined, only 1880 have an upper limit on their 95% credible interval that is greater than 0.1 mg/kg BW/day. These compounds are plotted in Panel a of Figure 4. The credible interval reflects a combination of scenarios where the values for predictors are large and the confidence in those predictors is significant. For the very highest predicted intake rates in Panel a, the credible intervals are very large, reflecting large uncertainty (exposure could be very high or very low). The prediction for the highest chemical exposure, Dihexyl nonanedioate, is driven by dietary exposure predicted by the Biyol et al. [46] model which, lacking any measured data, assumed that this compound might be a significant fraction of some packaging materials. Better characterization of the actual occurrence and weight fraction of chemicals such as Dihexyl nonanedioate, where assumptions may be too conservative, would reduce this uncertainty. The median credible interval in Panel a spans 8.5 orders of magnitude, with a high of 12.7 and a low of 4.5.

As shown in Panel b of Figure 4, there is 95% confidence that the median intake rate is below 0.1 µg/kg BW/day for 478,046 compounds. Chemicals produced/imported in quantities of more than 25,000 lbs./year ("high production volume chemicals") must be reported to the EPA, however, since production/import of most of the chemicals considered here are not listed as high production volume the actual production is unknown and we assume a default of 12,500 lbs./year. If these chemicals are instead produced at just below the reporting requirement, the number of chemicals exceeding 0.1 mg/kg BW/day would be 1962. The median credible interval in Panel b spans 4.3 orders of magnitude, with a high of 16.1 and a low of 0.9.

An exposure pathway may cover multiple “exposure routes”, defined as “The way an agent enters a target after contact (e.g., by ingestion, inhalation, or dermal absorption).”[14] Many of the models analyzed here can produce route-specific estimates, and the coarse, source-based pathway categories used here could be refined in the future if more data become available to predict likely routes for new chemicals (e.g. ingestion of industrial chemicals via water vs. inhalation of contaminated air). We believe that our current source-based pathway categories are an appropriate level of refinement given the level of information available for tens of thousands of chemicals.

The SEEM meta-model extrapolates intake rates from 114 chemicals that can be inferred from NHANES biomonitoring data. We make a strong assumption that chemicals without biomonitoring data will have similar intake rates to those within NHANES. The selection of NHANES chemicals by the CDC is not intended to represent all chemical space but does contain a mix of chemicals above and below the limits of detection [20]. While the NHANES chemicals do cover all four exposure pathways modeled here, there are certainly chemicals of interest in regions of chemical space (e.g., hydrophobicity, ionization, function) as well as exposure pathways that are not covered.

There are at least two reasons to desire expanding the available set of chemicals with biomonitoring data suitable for model evaluation. First, many chemicals with the highest predicted intake rates have not been included in targeted bio-monitoring efforts. New data on these chemicals would provide critical evaluation data. Second, statistical ability to evaluate model predictions generally increases with the number of samples and would be especially improved by expanding the chemical space covered.

The 95% credible intervals reported here reflect uncertainty about the population median intake rate, and do not reflect population variability. The NHANES data do not necessarily capture the most highly exposed individuals, including those with occupational exposures. There are some significant limitations for estimating intake rates for highly exposed individuals because any given chemical is monitored via spot samples for a cohort of roughly 2500 individuals[78]. Since a log-normal population distribution was assumed, the median (mean of the distribution) is more certain than the overall shape of the distribution. While spot samples are informative about median intake rates, variation in exposure magnitudes, duration, and time from sampling confounds the estimation of higher moments of the distribution [79].

The steady-state approximation for exposure reconstruction is less than ideal, particularly for chemicals with short half-lives and irregular use patterns [20, 80, 81]. At a minimum, data for chemical volume of distribution and clearance rate are needed to make more elaborate inferences [82, 83]. Unfortunately, these data are not available for all NHANES chemicals [38]. Despite these limitations, the median intake rate can be reasonably estimated from population biomarker data since those samples average over the various exposure scenarios – even for rapidly cleared chemicals [84]. Further, it has been shown via simulation study that most chemicals with potential environmental exposure do reach steady-state within a few weeks [85].

The biomarkers used here for evaluation and calibration represent the general U.S. population as characterized by NHANES. Separate analyses are possible for specific demographic groups (e.g., children, women of reproductive age) [20] if there is sufficient representation within the NHANES cohort but have not been performed here. We have not evaluated how representative NHANES is of non-U.S. populations, although the SEEM approach could be applied to any similar chemical exposure biomonitoring data set representing other. Similarly, the predictors examined here are primarily for the general U.S. population, but other exposure data and model scenarios could be used in consensus predictions for other populations.

Of 687,359 chemicals evaluated, 30% have less than a 50% probability for exposure via any of the four pathways modeled here. Since the various pathways act to either raise or lower predicted intake rate, these chemicals were predicted to have a moderate exposure rate (the grand mean exposure rate, $a_0$, from Equation 1). However, these chemicals should be thought of as being outside the "applicability domain"[17] of the meta-model because we have not characterized the uncertainty since there are no NHANES data for these chemicals. It is possible that these chemicals have no significant use resulting in exposure to the population, or that exposure to them occurs by a route different than those modeled here (e.g., pharmaceuticals). The source-based pathways used in this study – far-field pesticide, far-field industrial, dietary, and consumer – are imperfect and incomplete [86, 87]. These pathways are demonstrative, and could be refined and replaced with other, more descriptive pathways assuming there are sufficient data. Despite these limitations, the nearly two thousand chemicals in Figure 4a remain priorities since they are predicted to be like NHANES chemicals with high exposure.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENT

This project was supported by appointments to the Internship/Research Participation Program at the Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA. The authors thank Drs. Richard Judson and Kathie Dionisio for their helpful EPA technical reviews of the manuscript.

Funding Sources

The United States Environmental Protection Agency, through its Office of Research and Development, funded and managed the research described here. However, it may not necessarily reflect official Agency policy, and reference to commercial products or services does not constitute endorsement.

## ABBREVIATIONS

| | |
|---|---|
| **BW** | body weight |
| **CDC** | U.S. Centers for Disease Control and Prevention |
| **CDR** | EPA Chemical Data Reporting |
| **CPDat** | EPA Chemical and Products Database |

| | |
|---|---|
| **DSSTox** | Distributed Structure-Searchable Toxicity |
| **EPA** | U.S. Environmental Protection Agency |
| **ExpoCast** | the EPA Exposure Forecaster research program |
| **FDA** | U.S. Food and Drug Administration |
| **FINE** | Fugacity-based INdoor Exposure model |
| **HTE** | High Throughput Exposure models |
| **MCMC** | Markov Chain Monte Carlo |
| **NASEM** | U.S. National Academy of Sciences, Engineering, and Medicine |
| **NHANES** | CDC National Health and Nutrition Examination Survey |
| **OOB** | Random Forest Out of Bag error |
| **NTA** | non-targeted chemical analysis |
| **OPERA** | OPEn structure–activity/property Relationship App |
| **RAIDAR** | Risk Assessment IDentification And Ranking model |
| **RAIDAR-ICE** | RAIDAR-Indoor and Consumer Exposure model |
| **SEEM** | the Systematic Empirical Evaluation of Models framework |
| **RED** | EPA Pesticide Re-Registration Eligibility Document |
| **SHEDS-HT** | high-throughput Stochastic Human Exposure and Dose Simulation model |
| **SSA** | suspect screening chemical analysis |
| **TSCA** | Toxic Substances Control Act |
| **USEtox** | United Nations Environment Programme (UNEP) and Society for Environmental Toxicology and Chemistry (SETAC) toxicity and ecotoxicity characterization model |

## REFERENCES

1. Thomas RS; Philbert MA; Auerbach SS; Wetmore BA; Devito MJ; Cote I; Rowlands JC; Whelan MP; Hays SM; Andersen ME, Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework. Toxicological Sciences 2013, 136, (1), 4–18. [PubMed: 23958734]
2. National Research Council, Toxicity testing in the 21st century: A vision and a strategy. National Academies Press: 2007.
3. Dong Z; Liu Y; Duan L; Bekele D; Naidu R, Uncertainties in human health risk assessment of environmental contaminants: A review and perspective. Environ Int 2015, 85, 120–32. [PubMed: 26386465]

4. Judson R; Richard A; Dix DJ; Houck K; Martin M; Kavlock R; Dellarco V; Henry T; Holderman T; Sayre P, The toxicity data landscape for environmental chemicals. Environmental health perspectives 2009, 117, (5), 685. [PubMed: 19479008]

5. National Research Council Exposure Science in the 21st Century: A Vision and a Strategy; 2012.

6. Egeghy PP; Judson R; Gangwal S; Mosher S; Smith D; Vail J; Cohen Hubal EA, The exposure data landscape for manufactured chemicals. The Science of the total environment 2012, 414, 159–66. [PubMed: 22104386]

7. Kavlock RJ; Bahadori T; Barton-Maclaren TS; Gwinn MR; Rasenberg M; Thomas RS, Accelerating the Pace of Chemical Risk Assessment. Chemical Research in Toxicology 2018, 31, (5), 287–290. [PubMed: 29600706]

8. Wetmore BA; Wambaugh JF; Allen B; Ferguson SS; Sochaski MA; Setzer RW; Houck KA; Strope CL; Cantwell K; Judson RS; LeCluyse E; Clewell HJ; Thomas RS; Andersen ME, Incorporating High-Throughput Exposure Predictions With Dosimetry-Adjusted In Vitro Bioactivity to Inform Chemical Toxicity Testing. Toxicological Sciences 2015, 148, (1), 121–36. [PubMed: 26251325]

9. Fryer M; Collins CD; Ferrier H; Colvile RN; Nieuwenhuijsen MJ, Human exposure modelling for chemical risk assessment: a review of current approaches and research and policy implications. Environmental Science & Policy 2006, 9, (3), 261–274.

10. Mitchell J; Arnot JA; Jolliet O; Georgopoulos PG; Isukapalli S; Dasgupta S; Pandian M; Wambaugh J; Egeghy P; Hubal EAC, Comparison of modeling approaches to prioritize chemicals based on estimates of exposure and exposure potential. Science of the Total Environment 2013, 458, 555–567. [PubMed: 23707726]

11. Cohen Hubal EA; Richard A; Aylward L; Edwards S; Gallagher J; Goldsmith M-R; Isukapalli S; Tornero-Velez R; Weber E; Kavlock R, Advancing exposure characterization for chemical evaluation and risk assessment. Journal of Toxicology and Environmental Health, Part B 2010, 13, (2–4), 299–313.

12. Hubal EAC, Biologically relevant exposure science for 21st century toxicity testing. Toxicological sciences 2009, 111, (2), 226–232. [PubMed: 19602574]

13. Egeghy PP; Sheldon LS; Isaacs KK; Özkaynak H; Goldsmith M-R; Wambaugh JF; Judson RS; Buckley TJ, Computational exposure science: An emerging discipline to support 21st-century risk assessment. Environmental health perspectives 2016, 124, (6), 697. [PubMed: 26545029]

14. Zartarian V; Bahadori T; McKone T, Adoption of an official ISEA glossary. Journal of Exposure Analysis & Environmental Epidemiology 2005, 15, (1).

15. Clark K; Cousins IT; Mackay D, Assessment of critical exposure pathways In Series Anthropogenic Compounds, Springer: 2003; pp 227–262.

16. MacLeod M; Scheringer M; McKone TE; Hungerbuhler K, The state of multimedia mass-balance modeling in environmental science and decision-making. In ACS Publications: 2010.

17. Sushko I; Novotarskyi S; Körner R; Pandey AK; Cherkasov A; Li J; Gramatica P; Hansen K; Schroeter T; Müller K-R, Applicability domains for classification problems: benchmarking of distance to models for Ames mutagenicity set. Journal of chemical information and modeling 2010, 50, (12), 2094–2111. [PubMed: 21033656]

18. Chadeau-Hyam M; Athersuch TJ; Keun HC; De Iorio M; Ebbels TM; Jenab M; Sacerdote C; Bruce SJ; Holmes E; Vineis P, Meeting-in-the-middle using metabolic profiling–a strategy for the identification of intermediate biomarkers in cohort studies. Biomarkers 2011, 16, (1), 83–88. [PubMed: 21114379]

19. CDC National Health and Nutrition Examination Survey. http://www.cdc.gov/nchs/nhanes.htm

20. Wambaugh JF; Wang A; Dionisio KL; Frame A; Egeghy P; Judson R; Setzer RW, High throughput heuristics for prioritizing human exposure to environmental chemicals. Environmental Science and Technology 2014, 48, (21), 12760–7. [PubMed: 25343693]

21. Wambaugh JF; Setzer RW; Reif DM; Gangwal S; Mitchell-Blackwood J; Arnot JA; Joliet O; Frame A; Rabinowitz J; Knudsen TB; Judson RS; Egeghy P; Vallero D; Cohen Hubal EA, High-throughput models for exposure-based chemical prioritization in the ExpoCast project. Environmental Science and Technology 2013, 47, (15), 8479–88. [PubMed: 23758710]

22. Ring CL; Pearce RG; Setzer RW; Wetmore BA; Wambaugh JF, Identifying populations sensitive to environmental chemicals by simulating toxicokinetic variability. Environment International 2017, 106, 105–118. [PubMed: 28628784]

23. Katritzky AR; Lobanov VS; Karelson M, QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chemical Society Reviews 1995, 24, (4), 279–287.

24. R Core Team R: A language and environment for statistical computing, R Foundation for Statistical Computing: Vienna, Austria, 2016.

25. Williams AJ; Grulke CM; Edwards J; McEachran AD; Mansouri K; Baker NC; Patlewicz G; Shah I; Wambaugh JF; Judson RS, The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. Journal of Cheminformatics 2017, 9, (1), 61. [PubMed: 29185060]

26. Gelman A; Stern HS; Carlin JB; Dunson DB; Vehtari A; Rubin DB, Bayesian data analysis. Chapman and Hall/CRC: 2013.

27. Plummer M In JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, Proceedings of the 3rd international workshop on distributed statistical computing, 2003; Vienna: 2003; p 125.

28. Denwood MJ, runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. Journal of Statistical Software 2016, 71, (9), 1–25.

29. Breiman L, Random forests. Machine learning 2001, 45, (1), 5–32.

30. Liaw A; Wiener M, Classification and regression by randomForest. R news 2002, 2, (3), 18–22.

31. Lemon SC; Roy J; Clark MA; Friedmann PD; Rakowski W, Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of behavioral medicine 2003, 26, (3), 172–181. [PubMed: 14644693]

32. Hilal S; Karickhoff S; Carreira L, A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pKas. Quantitative Structure‐Activity Relationships 1995, 14, (4), 348–355.

33. Archer KJ; Kimes RV, Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis 2008, 52, (4), 2249–2260.

34. Mansouri K; Grulke CM; Judson RS; Williams AJ, OPERA models for predicting physicochemical properties and environmental fate endpoints. Journal of cheminformatics 2018, 10, (1), 10. [PubMed: 29520515]

35. Richard AM; Williams CR, Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 2002, 499, (1), 27–52. [PubMed: 11804603]

36. Yang C; Tarkhov A; Marusczyk J. r.; Bienfait B; Gasteiger J; Kleinoeder T; Magdziarz T; Sacher O; Schwab CH; Schwoebel J, New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. Journal of chemical information and modeling 2015, 55, (3), 510–528. [PubMed: 25647539]

37. Yap CW, PaDEL‐descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of computational chemistry 2011, 32, (7), 1466–1474. [PubMed: 21425294]

38. Pearce RG; Setzer RW; Strope CL; Sipes NS; Wambaugh JF, Httk: R package for high-throughput toxicokinetics. Journal of Statistical Software 2017, 79, (1), 1–26. [PubMed: 30220889]

39. Rotroff DM; Wetmore BA; Dix DJ; Ferguson SS; Clewell HJ; Houck KA; LeCluyse EL; Andersen ME; Judson RS; Smith CM, Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. Toxicological Sciences 2010, 117, (2), 348–358. [PubMed: 20639261]

40. Wetmore BA; Wambaugh JF; Ferguson SS; Sochaski MA; Rotroff DM; Freeman K; Clewell HJ 3rd; Dix DJ; Andersen ME; Houck KA; Allen B; Judson RS; Singh R; Kavlock RJ; Richard AM; Thomas RS, Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. Toxicological Sciences 2012, 125, (1), 157–74. [PubMed: 21948869]

41. Arnot JA; Brown TN; Wania F, Estimating screening-level organic chemical half-lives in humans. Environmental Science & Technology 2013, 48, (1), 723–730. [PubMed: 24298879]

42. U.S. Environmental Protection Agency, PFOS Health Advisory. In 2016.

43. U.S. Environmental Protection Agency, PFOA Health Advisory. In 2016.

44. U.S. Environmental Protection Agency Downloadable 2006 IUR Public Database. https://www.epa.gov/chemical-data-reporting/downloadable-2006-iur-public-database (5/2/17),

45. Lallas PL, The Stockholm Convention on persistent organic pollutants. American Journal of International Law 2001, 95, (3), 692–708.

46. Biryol D; Nicolas CI; Wambaugh J; Phillips K; Isaacs K, High-throughput dietary exposure predictions for chemical migrants from food contact substances for use in chemical prioritization. Environment International 2017, 108, 185–194. [PubMed: 28865378]

47. Administration, U. S. F. D. U.S. FDA. Guidance for Industry: Preparation of Premarket Submissions for Food Contact Substances. https://www.fda.gov/Food/GuidanceRegulation/GuidanceDocumentsRegulatoryInformation/IngredientsAdditivesGRASPackaging/ucm081818.htm#iid4

48. Rosenbaum RK; Bachmann TM; Gold LS; Huijbregts MA; Jolliet O; Juraske R; Koehler A; Larsen HF; MacLeod M; Margni M, USEtox—the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment. The International Journal of Life Cycle Assessment 2008, 13, (7), 532–546.

49. Wannaz C; Fantke P; Jolliet O, Multiscale Spatial Modeling of Human Exposure from Local Sources to Global Intake. Environmental Science & Technology 2018, 52, (2), 701–711. [PubMed: 29249158]

50. Fantke P; Juraske R; Antón A; Friedrich R; Jolliet O, Dynamic Multicrop Model to Characterize Impacts of Pesticides in Food. Environmental Science & Technology 2011, 45, (20), 8842–8849. [PubMed: 21905656]

51. Fantke P; Jolliet O, Life cycle human health impacts of 875 pesticides. The International Journal of Life Cycle Assessment 2016, 21, (5), 722–733.

52. Fantke P; Wieland P; Juraske R; Shaddick G; Itoiz ES; Friedrich R; Jolliet O, Parameterization Models for Pesticide Exposure via Crop Consumption. Environmental Science & Technology 2012, 46, (23), 12864–12872. [PubMed: 23136826]

53. Fantke P; Wieland P; Wannaz C; Friedrich R; Jolliet O, Dynamics of pesticide uptake into plants: From system functioning to parsimonious modeling. Environmental Modelling & Software 2013, 40, 316–324.

54. Arnot JA; Mackay D, Policies for chemical hazard and risk priority setting: can persistence, bioaccumulation, toxicity, and quantity information be combined? In ACS Publications: 2008.

55. Dionisio K; Phillips K; Price P; Grulke C; Williams A; Biryol D; Hong T; Isaacs K, The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. Scientific Data 2018, 5, 180125. [PubMed: 29989593]

56. Isaacs KK; Glen WG; Egeghy P; Goldsmith M-R; Smith L; Vallero D; Brooks R; Grulke CM; Özkaynak H. k., SHEDS-HT: an integrated probabilistic exposure model for prioritizing exposures to chemicals with near-field and dietary sources. Environmental Science & Technology 2014, 48, (21), 12750–12759. [PubMed: 25222184]

57. Zartarian V; Xue J; Glen G; Smith L; Tulve N; Tornero-Velez R, Quantifying children's aggregate (dietary and residential) exposure and dose to permethrin: application and evaluation of EPA's probabilistic SHEDS-Multimedia model. Journal of Exposure Science and Environmental Epidemiology 2012, 22, (3), 267–273. [PubMed: 22434114]

58. Bennett DH; Furtaw EJ, Fugacity-based indoor residential pesticide fate model. Environmental Science & Technology 2004, 38, (7), 2142–2152. [PubMed: 15112818]

59. Shin H-M; McKone TE; Bennett DH, Intake Fraction for the Indoor Environment: A Tool for Prioritizing Indoor Chemical Sources. Environmental Science & Technology 2012, 46, (18), 10063–10072. [PubMed: 22920860]

60. USEPA Exposure Factors Handbook 2011 Edition (Final); U.S. Environmental Protection Agency: Washington, DC, 2011.

61. Li L; Westgate JN; Hughes L; Zhang X; Givehchi B; Toose L; Armitage JM; Wania F; Egeghy P; Arnot JA, A model for risk-based screening and prioritization of human exposure to chemicals from near-field sources. Environmental science & technology 2018, in press.

62. Zhang X; Arnot JA; Wania F, Model for screening-level assessment of near-field human exposure to neutral organic chemicals released indoors. Environmental Science & Technology 2014, 48, (20), 12312–12319. [PubMed: 25264817]

63. Jolliet O; Ernstoff AS; Csiszar SA; Fantke P, Defining Product Intake Fraction to Quantify and Compare Exposure to Consumer Products. Environmental Science & Technology 2015, 49, (15), 8924–8931. [PubMed: 26102159]

64. Huang L; Ernstoff A; Fantke P; Csiszar SA; Jolliet O, A review of models for near-field exposure pathways of chemicals in consumer products. Science of The Total Environment 2017, 574, 1182–1208. [PubMed: 27644856]

65. Huang L; Jolliet O, A parsimonious model for the release of volatile organic compounds (VOCs) encapsulated in products. Atmospheric Environment 2016, 127, 223–235.

66. Ernstoff AS; Fantke P; Huang L; Jolliet O, High-throughput migration modelling for estimating exposure to chemicals in food packaging in screening and prioritization tools. Food and Chemical Toxicology 2017, 109, 428–438. [PubMed: 28939300]

67. Csiszar SA; Ernstoff AS; Fantke P; Meyer DE; Jolliet O, High-throughput exposure modeling to support prioritization of chemicals in personal care products. Chemosphere 2016, 163, 490–498. [PubMed: 27565317]

68. European Chemicals Agency (ECHA), New Approach Methodologies in Regulatory Science, Proceedings of a scientific workshop In Helsinki, 2016.

69. Shin H-M; Ernstoff AS; Arnot JA; Wetmore BA; Csiszar SA; Fantke P; Zhang X; McKone TE; Jolliet O; Bennett D, Risk-based high throughput chemical screening and prioritization using exposure models and in vitro bioacitivity assays. Environmental Science and Technology 2015, 49, 6760–6771. [PubMed: 25932772]

70. Fantke P; Ernstoff AS; Huang L; Csiszar SA; Jolliet O, Coupled near-field and far-field exposure assessment framework for chemicals in consumer products. Environment International 2016, 94, 508–518. [PubMed: 27318619]

71. Rager JE; Strynar MJ; Liang S; McMahen RL; Richard AM; Grulke CM; Wambaugh JF; Isaacs KK; Judson R; Williams AJ, Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. Environment International 2016, 88, 269–280. [PubMed: 26812473]

72. Park YH; Lee K; Soltow QA; Strobel FH; Brigham KL; Parker RE; Wilson ME; Sutliff RL; Mansfield KG; Wachtman LM; Ziegler TR; Jones DP, High-performance metabolic profiling of plasma from seven mammalian species for simultaneous environmental chemical surveillance and bioeffect monitoring. Toxicology 2012, 295, (1–3), 47–55. [PubMed: 22387982]

73. Phillips K; Yau AY; Favela KA; Isaacs KK; McEachran A; Grulke CM; Richard AM; Williams A; Sobus JR; Thomas RS; Wambaugh JF, Suspect screening analysis of chemicals in consumer products. Environmental Science & Technology 2018, 52, (5), 3125–3135. [PubMed: 29405058]

74. Phillips KA; Wambaugh JF; Grulke CM; Dionisio KL; Isaacs KK, High-throughput screening of chemicals as functional substitutes using structure-based classification models. Green Chemistry 2017, 19, (4), 1063–1074. [PubMed: 30505234]

75. Boxall A; Keller V; Straub J; Monteiro S; Fussell R; Williams R, Exploiting monitoring data in environmental exposure modelling and risk assessment of pharmaceuticals. Environment international 2014, 73, 176–185. [PubMed: 25127044]

76. Schymanski EL; Singer HP; Slobodnik J; Ipolyi IM; Oswald P; Krauss M; Schulze T; Haglund P; Letzel T; Grosse S, Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. Analytical and bioanalytical chemistry 2015, 407, (21), 6237–6255. [PubMed: 25976391]

77. Li L; Arnot J; Wania F, Revisiting the Contributions of Far-and Near-Field Routes to Aggregate Human Exposure to Polychlorinated Biphenyls (PCBs). Environmental science & technology 2018, 52, (12), 6974–6984. [PubMed: 29771504]

78. Kapraun DF; Wambaugh JF; Ring CL; Tornero-Velez R; Setzer RW, A Method for Identifying Prevalent Chemical Combinations in the US Population. Environmental Health Perspectives 2017, 125, (8), 087017. [PubMed: 28858827]

79. Aylward LL; Kirman CR; Schoeny R; Portier CJ; Hays SM, Evaluation of biomonitoring data from the CDC National Exposure Report in a risk assessment context: perspectives across chemicals. Environmental health perspectives 2013, 121, (3), 287. [PubMed: 23232556]

80. Hays SM; Becker RA; Leung H-W; Aylward LL; Pyatt DW, Biomonitoring equivalents: a screening approach for interpreting biomonitoring results from a public health risk perspective. Regulatory Toxicology and Pharmacology 2007, 47, (1), 96–109. [PubMed: 17030369]

81. Aylward LL; Kirman CR; Adgate JL; McKenzie LM; Hays SM, Interpreting variability in population biomonitoring data: role of elimination kinetics. Journal of Exposure Science and Environmental Epidemiology 2012, 22, (4), 398. [PubMed: 22588214]

82. Georgopoulos PG; Sasso AF; Isukapalli SS; Lioy PJ; Vallero DA; Okino M; Reiter L, Reconstructing population exposures to environmental chemicals from biomarkers: challenges and opportunities. Journal of Exposure Science and Environmental Epidemiology 2009, 19, (2), 149. [PubMed: 18368010]

83. Sobus JR; Tan Y-M; Pleil JD; Sheldon LS, A biomonitoring framework to support exposure and risk assessments. Science of the Total Environment 2011, 409, (22), 4875–4884. [PubMed: 21906784]

84. Aylward LL; Hays SM; Zidek A, Variation in urinary spot sample, 24 h samples, and longer-term average urinary concentrations of short-lived environmental chemicals: implications for exposure assessment and reverse dosimetry. Journal of Exposure Science and Environmental Epidemiology 2017, 27, (6), 582. [PubMed: 27703149]

85. Wambaugh JF; Wetmore BA; Pearce R; Strope C; Goldsmith R; Sluka JP; Sedykh A; Tropsha A; Bosgra S; Shah I; Judson R; Thomas RS; Setzer RW, Toxicokinetic Triage for Environmental Chemicals. Toxicological Sciences 2015, 147, (1), 55–67. [PubMed: 26085347]

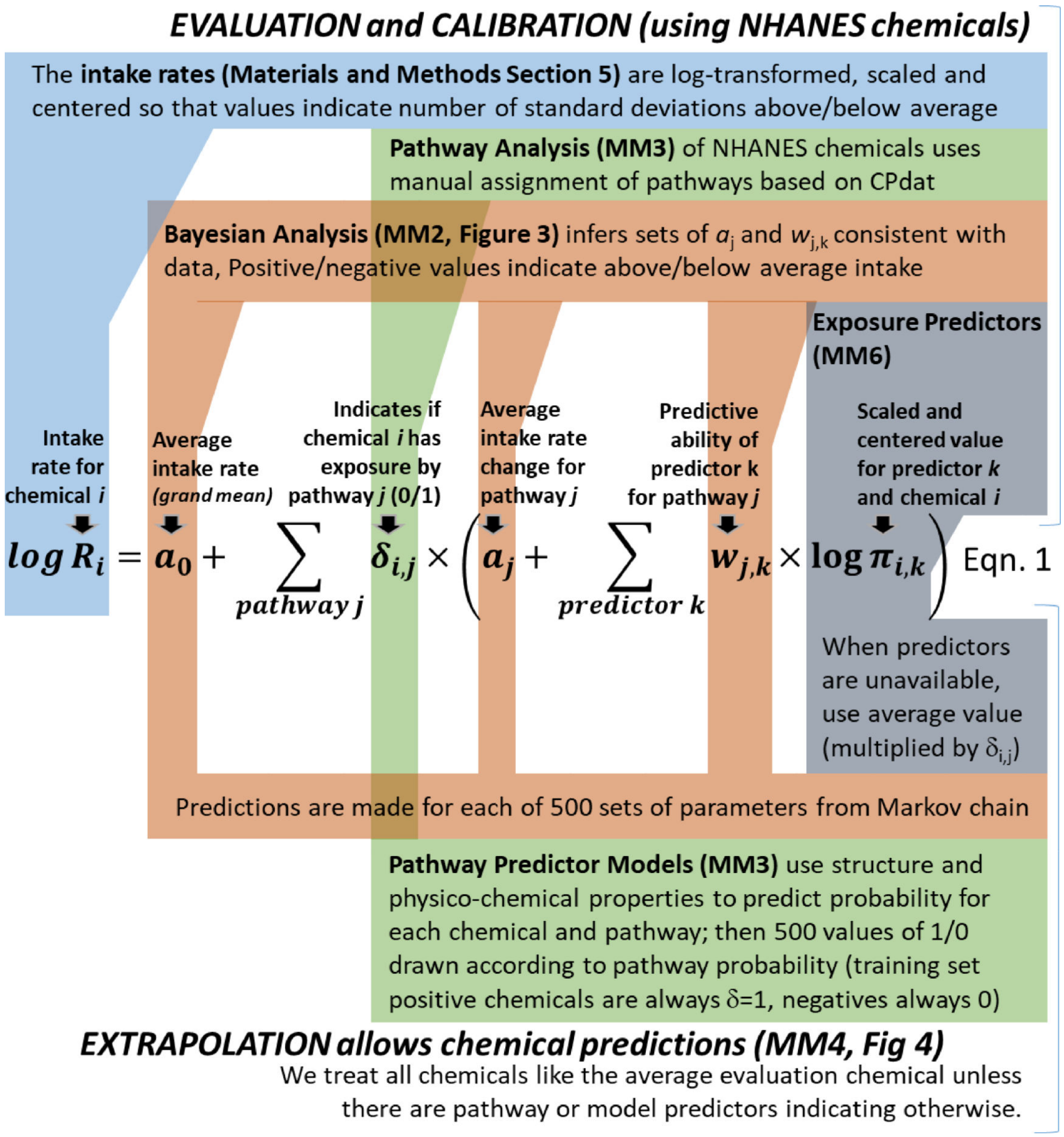
86. Arnot JA; MacKay D; Webster E; Southwood JM, Screening level risk assessment model for chemical fate and effects in the environment. Environmental Science & Technology 2006, 40, (7), 2316–2323. [PubMed: 16646468]

87. Mattingly CJ; McKone TE; Callahan MA; Blake JA; Hubal EAC, Providing the Missing Link: the Exposure Science Ontology ExO. Environmental Science & Technology 2012, 46, (6), 3046–3053. [PubMed: 22324457]

88. Sheldon LS; Cohen Hubal EA, Exposure as part of a systems approach for assessing risk. Environmental Health Perspectives 2009, 117, (8), 1181.

89. Everitt BS, Medical statistics from A to Z: a guide for clinicians and medical students. Cambridge University Press: 2006.

90. Isaacs K R Package SHEDS-HT, 2017.



EVALUATION and CALIBRATION (using NHANES chemicals)
Exposure Predictors (MM6)
Intake rate for chemical *i*
Average intake rate *(grand mean)*
Indicates if chemical *i* has exposure by pathway *j* (0/1)
Average intake rate change for pathway *j*
Predictive ability of predictor k for pathway *j*
Scaled and centered value for predictor *k* and chemical *i*
EXTRAPOLATION allows chemical predictions (MM4, Fig 4)

The **intake rates (Materials and Methods Section 5)** are log-transformed, scaled and centered so that values indicate number of standard deviations above/below average

**Pathway Analysis (MM3)** of NHANES chemicals uses manual assignment of pathways based on CPdat

**Bayesian Analysis (MM2, Figure 3)** infers sets of $a_j$ and $w_{j,k}$ consistent with data, Positive/negative values indicate above/below average intake

$$\log R_i = a_0 + \sum_{pathway\ j} \delta_{i,j} \times \left( a_j + \sum_{predictor\ k} w_{j,k} \times \log \pi_{i,k} \right) \quad \text{Eqn. 1}$$

When predictors are unavailable, use average value (multiplied by $\delta_{i,j}$)

Predictions are made for each of 500 sets of parameters from Markov chain

**Pathway Predictor Models (MM3)** use structure and physico-chemical properties to predict probability for each chemical and pathway; then 500 values of 1/0 drawn according to pathway probability (training set positive chemicals are always $\delta$=1, negatives always 0)

We treat all chemicals like the average evaluation chemical unless there are pathway or model predictors indicating otherwise.

**Figure 1:**

The Systematic Empirical Evaluation of Models (SEEM) framework combines predictors of exposure ($\pi_{i,k}$) according to how well they explain the available intake rates ($R_i$). In the top half of the figure we describe how the overall average (grand mean) $a_0$, the pathway averages $a_j$, and the model weights $w_{j,k}$ are determined with Bayesian analysis. Each $w_{j,k}$ is an evaluation of each predictor, as well as a calibration of how to align that predictor with the intake rates. In the bottom half of the figure we explain how for chemicals without intake rates, we extrapolate the averages and weights from the Bayesian analysis to combine the predictors into a consensus prediction. The predictors are centered such that if no predictor is available, the average value is used. The pathway indicators $\delta_{i,j}$, are predicted using the Random Forest algorithm (Table 3).

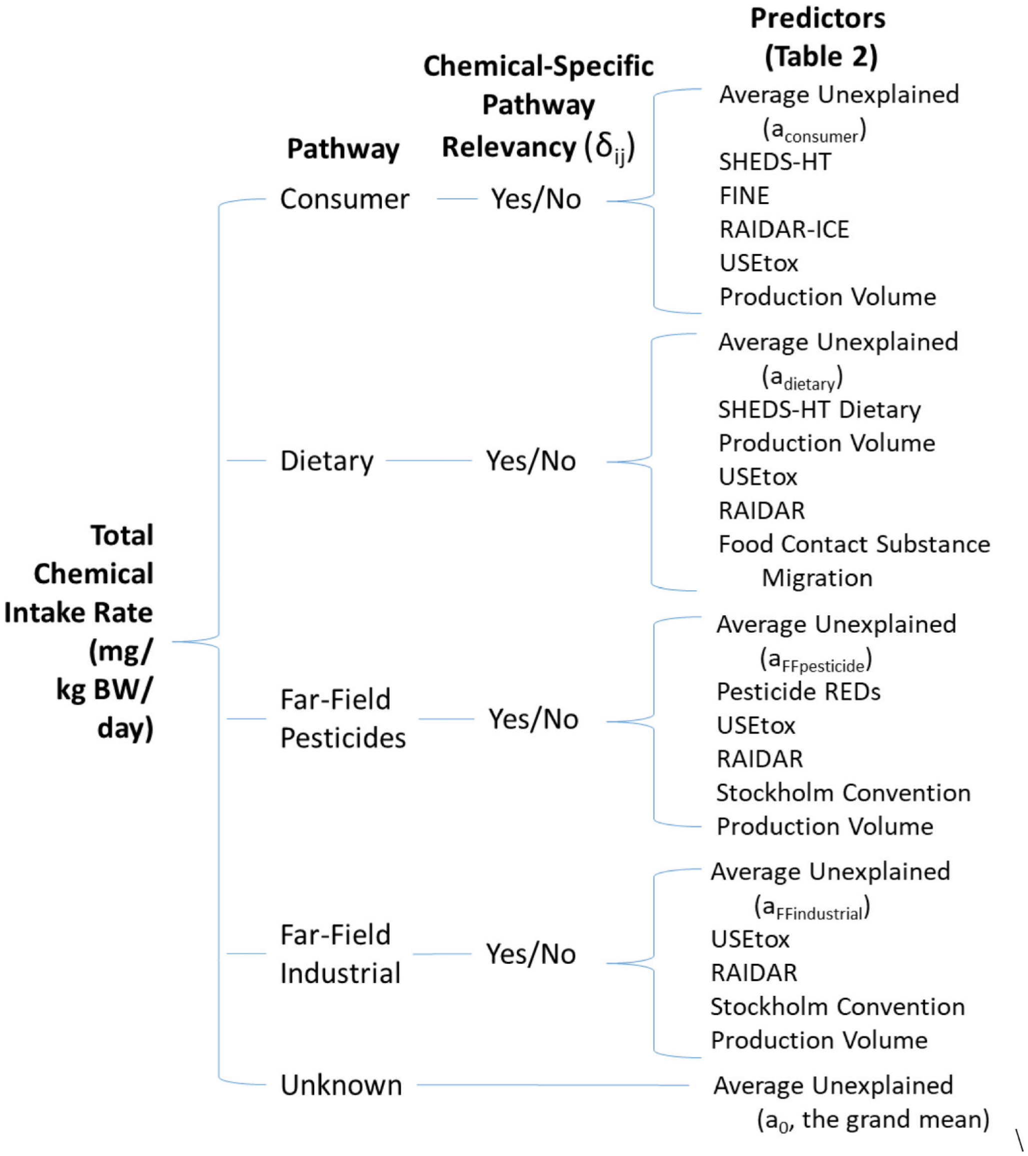

Predictors (Table 2)
Chemical-Specific Pathway Relevancy ($\delta_{ij}$)
Pathway
Total Chemical Intake Rate (mg/ kg BW/ day)
Consumer
Yes/No
Average Unexplained ($a_{consumer}$)
SHEDS-HT
FINE
RAIDAR-ICE
USEtox
Production Volume
Dietary
Yes/No
Average Unexplained ($a_{dietary}$)
SHEDS-HT Dietary
Production Volume
USEtox
RAIDAR
Food Contact Substance Migration
Far-Field Pesticides
Yes/No
Average Unexplained ($a_{FFpesticide}$)
Pesticide REDs
USEtox
RAIDAR
Stockholm Convention
Production Volume
Far-Field Industrial
Yes/No
Average Unexplained ($a_{FFindustrial}$)
USEtox
RAIDAR
Stockholm Convention
Production Volume
Unknown
Average Unexplained ($a_0$, the grand mean)
\

**Figure 2:**

Exposure predictors are organized to give a consensus prediction of intake rate based upon the exposure pathway(s) associated with a chemical. The exposure pathway indicators ($\delta_{ij}$) determine whether (1) or not (0) each pathway is associated --- if 1 (“yes”), then the predictors to the right will modify the estimated intake rate. The pathway means ($a_i$) indicate overall relative changes in intake rate associated with each pathway. Each exposure predictor and the NHANES intake rates were scaled so that their mean was zero and any value indicates the number of standard deviations above or below the mean. When a given predictor is unavailable for a given chemical, the mean value is used.

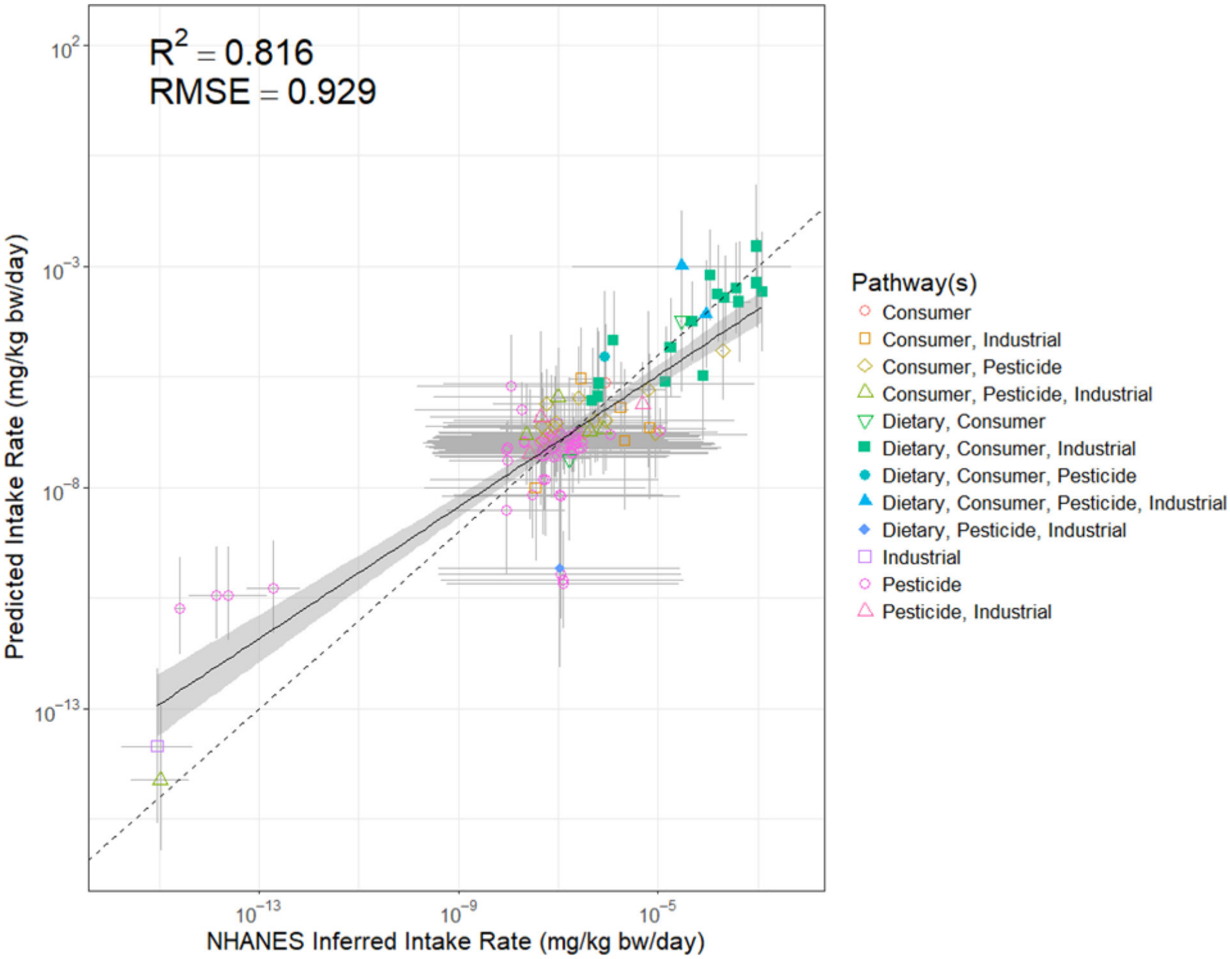


**Figure 3:**

A pathway-based SEEM meta-analysis allows disparate sources of exposure information (e.g., expert estimates of intake rate and high-throughput exposure estimates, both described here as “models”). Chemicals with no predictions for a given model are assigned either the average exposure predicted for that model or zero, depending on whether or not a chemical is predicted to have exposure via the pathway relevant to that model. Most of the 114 NHANES chemicals analyzed are predicted to have exposure via multiple pathways and are distributed according to Table 2. The unexplained chemical-to-chemical variability is an empirical estimate of the uncertainty of our calibrated predictions. The dashed line indicates identity (perfect predictor) while the solid line indicates a least squares regression on the medians (with gray shaded region indicating standard error).

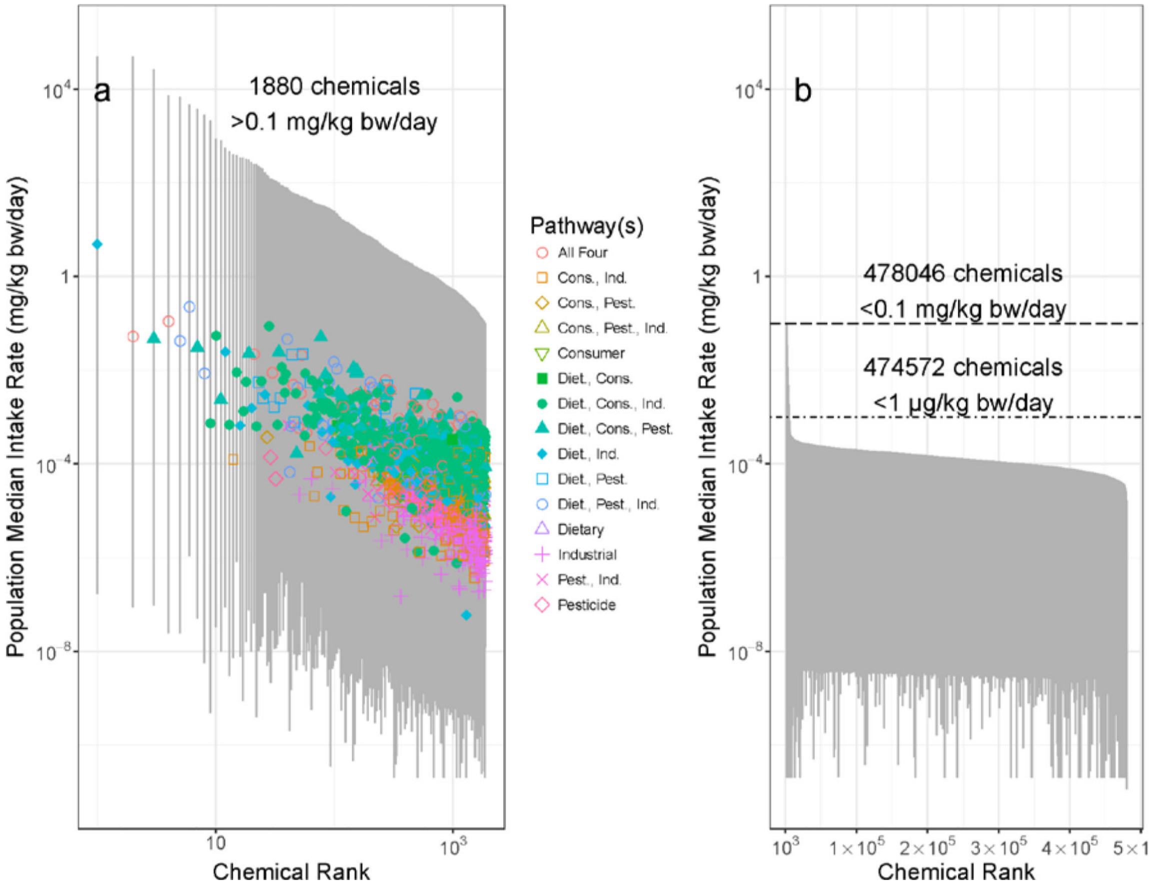


**Figure 4:**

The 95% credible interval (vertical line) and median predicted exposure (points in Panel A) for 479,926 chemicals. The 1880 chemicals whose confidence intervals exceed 0.1 mg/kg BW/day are ranked on a logarithmic scale (Panel a) and all remaining chemicals plotted on an arithmetic scale (panel b). The shape of each plot-point in Panel a indicates the predicted (>50% probability) or assumed (training set) exposure pathways. Chemicals may have exposure by none (i.e., “unknown” pathway), one, or more than one of the four pathways. The upper limit of the 95% interval for the vast majority of chemicals is less than 1 μg/kg BW/day. The upper limit of the credible interval for the first four chemicals in panel A is truncated for plotting clarity.

## Table 1:

Partial Technical Glossary

| Term | Explanation |
|---|---|
| **ExpoCast (Exposure Forecasting) Project** | An ongoing U.S. Environmental Protection Agency project to develop new methods, data, and models for high-throughput exposure assessment (i.e., thousands of chemicals). [11, 88] |
| **Exposure Predictor** | In this analysis "exposure predictor" refers to both the predictions of specific exposure models as well as other exposure-related information |
| **Exposure Pathway** | "The course an agent [chemical] takes from the source [environmental release] to the target [human]." [14] In this analysis we use the simple term "pathway" to represent the totality of paths that a chemical may follow from a particular source to reach a person. |
| **Grand Mean** | The overall mean of a regression. In this analysis, the grand mean $a_0$ describes the average intake rate inferred from NHANES in contrast to the pathway-specific means. [89] |
| **Intake** | "The process by which an agent [chemical] crosses an outer exposure surface [some portion of an individual] of a target [human] without passing an absorption barrier, i.e. through ingestion or inhalation." [14] |
| **Intake rate** | Daily average intake (mg/kg body weight/day). |
| **Meet-in-the-Middle** | An approach in which predictions from models that make predictions from upstream data (e.g., activity) are compared with models that make inferences from downstream data (e.g., biomarkers). An approach in which predictions from models that make predictions from upstream data (e.g., [18] |
| **Near-field / Far-field Sources** | ”Near-field" sources are proximate, indoor sources such as consumer product use in domestic settings, while "far-field" sources are distal with exposure mediated by environmental fate and transport. [56, 70, 86] |
| **Random Forest Algorithm** | A machine learning approach in which an ensemble of decision trees is used to make probabilistic predictions. [29] |
| **Systematic Empirical Evaluation of Models (SEEM)** | SEEM is a consensus modeling method for exposure model evaluation and calibration. SEEM uses a meet-in-the-middle approach to calibrate high-throughput exposure predictors with intake rates inferred from biomonitoring data.[20, 21] |

## Table 2:

Exposure Predictors Evaluated

| **Predictor** | **Materials and Methods Section** | **Chemicals Predicted** | **Pathways** |
|---|---|---|---|
| EPA Inventory Update Reporting and Chemical Data Reporting (CDR) (2015)[44]EPA Inventory Update Reporting and Chemical Data Reporting (CDR) (2015)[44] | MM6.1 | 7856 | All |
| Stockholm Convention of Banned Persistent Organic Pollutants (2017)[45]Stockholm Convention of Banned Persistent Organic Pollutants (2017)[45] | MM6.2 | 248 | Far-Field Industrial and Pesticide |
| EPA Pesticide Reregistration Eligibility Documents (REDs) Exposure Assessments (Through 2015)[8, 40] | MM6.3 | 239 | Far-Field Pesticide |
| Food Contact Substance Migration Model (2017)[46]Food Contact Substance Migration Model (2017)[46] | MM6.4 | 940 | Dietary |
| United Nations Environment Program and Society for Environmental Toxicology and Chemistry toxicity model (USEtox) Industrial Scenario (2.0)[50–53] | MM6.5.1 | 8167 | Far-Field Industrial |
| USEtox Pesticide Scenario (2.0)[48]USEtox Pesticide Scenario (2.0)[48] | MM6.5.2 | 8167 | Far-Field Pesticide |
| Risk Assessment IDentification And Ranking (RAIDAR) Far-Field (2.95)[54]Risk Assessment IDentification And Ranking (RAIDAR) Far-Field (2.95)[54] | MM6.5.3 | 7511 | Far-Field Industrial and Pesticide |
| EPA Stochastic Human Exposure Dose Simulator High-Throughput (SHEDS-HT) Near-Field Direct (2017)[90]EPA Stochastic Human Exposure Dose Simulator High-Throughput (SHEDS-HT) Near-Field Direct (2017)[90] | MM6.6.1 | 1119 | Consumer (Near-Field) |
| SHEDS-HT Near-field Indirect (2017)[90]SHEDS-HT Near-field Indirect (2017)[90] | MM6.6.1 | 645 | Consumer |
| Fugacity-based INdoor Exposure (FINE) (2017)[58, 59]Fugacity-based INdoor Exposure (FINE) (2017)[58, 59] | MM6.6.2 | 1221 | Consumer |
| RAIDAR-ICE Near-Field (0.804)[61, 62] RAIDAR-ICE Near-Field (0.804)[61, 62] | MM6.6.3 | 615 | Consumer |
| USEtox Consumer Scenario (2.0)[63–65] | MM6.6.4 | 8167 | Consumer |
| USEtox Dietary Scenario (2.0)[63, 65, 66] | MM6.6.4 | 8167 | Dietary |

**Table 3:**

Training Sets and Performance of Random Forest Models for Exposure Pathways

| | NHANES Chemicals | Positives | Negatives | OOB Error Rate | Positives Error Rate | Balanced Accuracy | Sources of Positive Example Chemicals | Sources of Negative Example Chemicals |
|---|---|---|---|---|---|---|---|---|
| **Dietary** | 24 | 2523 | 8865 | 27 | 32 | 73 | FDA CEDI, ExpoCast, CPDat (Food, Food Additive, Food Contact), NHANES Curation | Pharmapendium, CPDat (non-food), NHANES Curation |
| **Consumer** | 49 | 1622 | 567 | 26 | 24 | 74 | CPDat (consumer_use, building_material), ExpoCast, NHANES Curation | CPDat (Agricultural, Industrial), FDA CEDI, NHANES Curation |
| **Far-Field Pesticide Sources** | 94 | 1480 | 6522 | 21 | 36 | 80 | REDs, Swiss Pesticides, Stockholm Convention, CPDat (Pesticide), NHANES Curation | Pharmapendium, Industrial Positives, NHANES Curation |
| **Far Field Industrial Sources** | 42 | 5089 | 2913 | 19 | 16 | 81 | CDR HPV, USGS Water Occurrence, NORMAN PFAS, Stockholm Convention, CPDat (Industrial, Industrial_Fluid), NHANES Curation | Pharmapendium, Pesticide Positives, NHANES Curation |