Toxicology Reports 3 (2016) 723–732

Contents lists available at ScienceDirect

Toxicology Reports

journal homepage: www.elsevier.com/locate/toxrep

Characterization and prediction of chemical functions and weight fractions in consumer products

Kristin K. Isaacs[a,*], Michael-Rock Goldsmith[b,1], Peter Egeghy[a], Katherine Phillips[c], Raina Brooks[d], Tao Hong[e], John F. Wambaugh[f]

[a] U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[b] Chemical Computing Group, Suite 910, 1010 Sherbrooke Street West, Montreal, QC H3A 2R7, Canada
[c] Oak Ridge Institute for Science and Education, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[d] Student Services Contractor, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[e] ICF International, 2635 Meridian Pkwy #200, Durham, NC 27713, United States
[f] U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States

ARTICLE INFO

Article history:
Received 17 May 2016
Received in revised form 5 August 2016
Accepted 31 August 2016
Available online 1 September 2016

Keywords:
Chemical function
Exposure modeling
Chemical prioritization
Consumer products
Cosmetics
ExpoCast

ABSTRACT

Assessing exposures from the thousands of chemicals in commerce requires quantitative information on the chemical constituents of consumer products. Unfortunately, gaps in available composition data prevent assessment of exposure to chemicals in many products. Here we propose filling these gaps via consideration of chemical functional role. We obtained function information for thousands of chemicals from public sources and used a clustering algorithm to assign chemicals into 35 harmonized function categories (e.g., plasticizers, antimicrobials, solvents). We combined these functions with weight fraction data for 4115 personal care products (PCPs) to characterize the composition of 66 different product categories (e.g., shampoos). We analyzed the combined weight fraction/function dataset using machine learning techniques to develop quantitative structure property relationship (QSPR) classifier models for 22 functions and for weight fraction, based on chemical-specific descriptors (including chemical properties). We applied these classifier models to a library of 10196 data-poor chemicals. Our predictions of chemical function and composition will inform exposure-based screening of chemicals in PCPs for combination with hazard data in risk-based evaluation frameworks. As new information becomes available, this approach can be applied to other classes of products and the chemicals they contain in order to provide essential consumer product data for use in exposure-based chemical prioritization.

1. Introduction

Assessment of the risks associated with chemicals in consumer products relies not only on characterization of hazard or toxicity, but also on the exposures encountered during use [1,2]. Consumer products contain and can release large numbers of chemicals to which humans are exposed directly during use or indirectly via contact with contaminated household air or dust [3–9]. Consumer product chemicals have been widely found in human blood and urine, and exposures from indoor or proximate "near-field" sources (which include consumer formulations and articles) generally are larger than the doses that result from "far-field" (e.g., industrial) sources of exposure [10–12].

Despite such high potential for exposure, critical gaps exist in both qualitative information describing the variety of chemicals contained in different categories of consumer products and in quantitative data on the weight fractions, both of which are key inputs to numerous exposure assessment frameworks and models [13–18]. However, due to limited public reporting requirements, confidential business considerations, and lack of harmonized chemical and product categorizations, specific data describing the composition of consumer products are often unavailable or incomplete [19,20].

These critical data gaps impede the quantification of exposures due to consumer product sources, and are especially noteworthy when considered in the context of prioritizing thousands of untested commercial chemicals on the basis of risk. The U.S. Environmental Protection Agency (EPA), under its ExpoCast program, is developing high-throughput (HT) computational methods for prediction of chemical exposures for combination with *in vitro* hazard information [1], with a particular goal of developing expo-

* Corresponding author.
E-mail address: isaacs.kristin@epa.gov (K.K. Isaacs).
[1] Work performed while an EPA employee.

http://dx.doi.org/10.1016/j.toxrep.2016.08.011

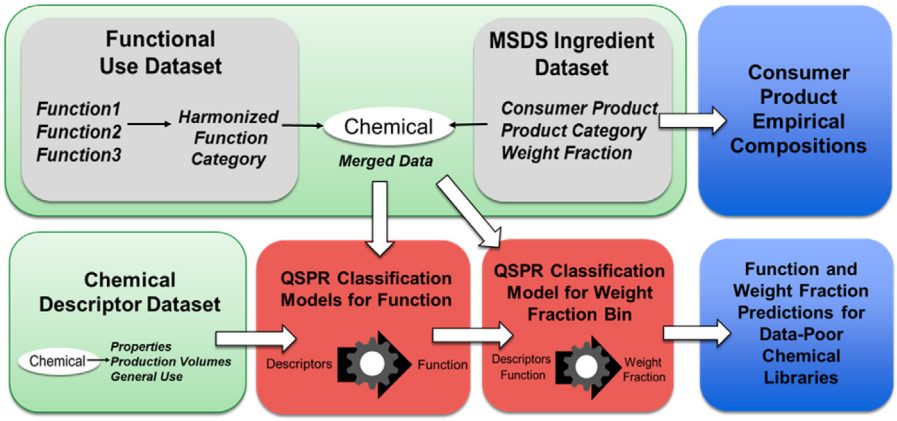

Fig. 1. Workflow for using existing chemical function and weight fraction datasets to build empirical compositions and QSPR classification models for chemical function and weight fraction for use in estimation of chemical exposure.

sure estimates for chemicals being evaluated by the ToxCast [21] initiative and the Tox21 interagency consortium [22]. A recent focus of ExpoCast has been the development of improved near-field (e.g., residential) exposures using both empirical [10,23] and mechanistic [13] approaches. To parameterize these efforts, U.S. EPA has developed new sources of information on how chemicals are used in commerce. EPA's Chemical/Product Categories Database (CPCat) [24,25] is a harmonized index of chemical use in products and sectors based on multiple publicly available data sources. One source within CPCat, the Consumer Product Chemical Profile Database (CPCPdb) [26] contains product ingredients and quantitative weight fractions derived from Material Safety Data Sheets (MSDS) for 1797 chemicals in nearly 9000 consumer products. Unfortunately, these quantitative data are limited to a relatively small fraction of products (and chemicals) currently in commerce. Methods are needed for extrapolating this existing knowledge to additional products and chemicals in a systematic manner.

In this work we present an approach for filling gaps in consumer product chemical use and composition data based on chemical function, and apply it to a case study of chemicals in personal care products (PCPs). Intentionally-added chemicals are present in consumer products because they serve a specific functional role that addresses either product performance or marketability. The functional role of an ingredient is defined by the chemical's properties and aids in determining its weight fraction in products. For example, Chevillotte et al. [27] described an exposure assessment method for cosmetics based on developing a "standard" or "virtual" composition of a product based on the weight fractions associated with chemical "families" across multiple product formulations. These families included functions such as "plasticizer" and "solvent." Here, we build on this approach by collecting and curating publicly-available function categorizations for thousands of chemicals, and combine these function categories with MSDS-based product weight fractions to build empirical compositions (or general formulations) based on real products in commerce for 66 categories of PCPs. These empirical compositions will be useful for parameterizing consumer exposure models for new or existing PCP chemicals when quantitative composition information is not available.

In addition to generating virtual compositions, we also describe a framework for predicting the probability of an arbitrary chemical having a given functional role and associated product weight fraction. This framework combines the function and ingredient weight fraction data to generate a series of machine-learning quantitative structure property relationship (QSPR) classification models for predicting functional role and weight fraction for large numbers of chemicals from chemical properties and other available descriptors (Fig. 1). These supervised learning models make use of known information about the characteristics of chemicals having certain functions to classify chemicals for which function is unknown. We apply these models to predict chemical functions for a library of over 10000 chemicals that are mostly data-poor, and corresponding weight fractions for hundreds of chemicals known to be present in PCPs. These methods are flexible and can be extended to additional chemical functions, products, or use sectors in support of HT prioritization of large numbers of chemicals on the basis of exposure potential or risk.

2. Methods

2.1. Chemical function data

Data describing the functions associated with individual chemicals (identified by Chemical Abstract Service Registry Numbers, CASRNs) were obtained from publicly-available government and industry sources; these data were curated into a harmonized Functional Use (FUse) database. Details (including sources) are provided in the Supplemental Information (SI). The largest source of data was the European Commission's Cosmetic Ingredient Database (CosIng) [28]. CosIng identifies different functional roles for cosmetic ingredients; a cosmetic is defined in CosIng to include a wide range of PCPs including lotions and creams, make-up, hair and body cleansing products, dental care products, fragrances, deodorants and antiperspirants, and suncreens [29]. A total of 10373 unique chemicals in PCPs were identified.

Many of the chemicals in the database were associated with multiple function categories. For the purpose of this study, we harmonized the function categories based on the similarity of the chemical groups associated with each category. For example, the majority of the chemicals classified as surfactants were also identified as cleansers and/or emulsifiers, so these chemicals were combined into a single harmonized category. This harmonization was based on a cluster analysis[30] of the function "fingerprint" of chemicals; a total of 36 harmonized functional categories relevant to PCPs were identified (details of the analysis and the results provided in the SI). A few chemicals had a greater variety of function classifications (e.g., "ethyl alcohol" had six functions: antifoamer, antimicrobial, astringent, solvent, masking agent, and viscosity controller). These chemicals were not assigned a single harmonized function for the purpose of calculating empirical compositions. Instead, these chemicals were categorized by name in the dataset, as they take on a variety of functions across different

product categories. These chemicals included zinc oxide, methyl salicylate, zinc pyrithione, ethyl alcohol, and sodium bicarbonate.

2.2. MSDS ingredient data

A database of MSDS-derived ingredient weight fractions in PCPs was developed for use in generating empirical compositions and predictive models. This database included 2433 PCPs reported in the CPCPdb [26], and an additional 1682 PCPs collected from product MSDS sheets provided online by manufacturers. Only ingredients for which CASRN were reported were retained. Details of the data sources and data collection are provided in the SI.

We previously assigned the products in this dataset to harmonized consumer product categories developed for exposure modeling purposes [13]. In that analysis, categories were aggregated to a specificity dictated by the available consumer product use information (e.g., population prevalence or frequency of use). For generation of the empirical compositions, however, the available PCP categories were refined where possible. For example, the category "eye makeup," was subdivided here into eyeshadows, eyeliners, and mascaras. In addition, several categories were further refined by form (e.g., gel, spray, powder) where indicated by the name of the product, and professional-use products (e.g., hair colors) were identified. A total of 66 categories were defined; final categories and number of products in each are listed in SI Table 2.

2.3. Function-based empirical compositions for personal care product categories

The harmonized chemical function (FUse) dataset (comprising unique function-CASRN pairs) and the ingredient weight fraction data were merged by CASRN for calculating empirical compositions for each product category. Summary statistics (mean, standard deviation, and select percentiles) for the weight fractions associated with each chemical function within each PCP category were calculated using the SAS UNIVARIATE procedure. Some MSDS reported nominal ranges (e.g. "0.1-1.0%"), in those cases, we chose to use the maximum of the range to be conservative. The number of unique chemicals and the chemical most frequently associated with each function for each product category were also determined, as was fraction of the products (or formulations) in each category containing at least one chemical with a given function.

2.4. Machine-learning models of function and weight fraction for use in chemical screening and prioritization

The merged function-ingredient dataset was used to develop a series of machine-learning QSPR classification models [31] for both function and weight fraction (Fig. 1). QSPR models describe the relationship between a chemical's known descriptors (e.g. structural or physiochemical information) and another property or characteristic of the chemical. QSPR models are based on either regression or classification methods, and can employ a variety of data-driven statistical techniques. The classification models built here take categorical or continuous chemical descriptors (i.e., predictive variables) as input and return assignment of the chemical into the class of interest (herein function or weight fraction bin). These descriptors (defined in SI Table S3) included 13 predicted or measured chemical properties obtained from EPI-Suite [32] and 16 simple descriptors of chemical use previously developed for the Tox21 chemical library and evaluated for inclusion in heuristic models of exposure [23]. Descriptors were available for 2981 chemicals for building the function models. Multiple classification models (one for each function with >10 chemicals for which descriptors were available) were built using random forests [33] with the R [34] package randomForest [35]. Random forest classifiers are ensembles of decision trees; each tree is built from a sampled subset of the test data. The classification models were built by analyzing the descriptors for all the chemicals that had a given function versus all the chemicals that did not; descriptors that best "separate" these two groups were identified. Each resulting model returns a probability of an arbitrary chemical performing the function based on its descriptors; this probability is equal to the fraction of the trees in the forest returning a positive classification for the chemical. Models were built using 5000 decision trees and downsampling [36] was implemented to account for imbalanced groups in the data. Estimates of the model error, sensitivity, specificity, and balanced accuracy (BA; mean of the specificity and sensitivity) were obtained using 5-fold cross-validation [37]. In addition, the method of y-scrambling [38] was used to further test the validity of the predictive models; models for each function were built for 10 sets of randomly-scrambled dependent variables (yes/no classifications for each function) and the mean and range of errors compared with the true model errors. Models with error greater than or equal to those generated by using the y-scrambled data were considered invalid.

An additional random forest model for weight fraction was built using a subset of the functional use dataset that could be merged with the ingredient weight fraction data; 17103 observations (828 chemicals) could be matched to the existing descriptors. The continuous quantitative weight fractions in the ingredient data were transformed using an logit (inverse logistic) function and then divided into three weight fraction bins (high: 0.3-1.0, medium: 0.01-0.3, and low: 0-0.01) for use in the predictive model; candidate bin boundaries were determined by a visual examination of a histogram of the transformed data (SI Fig. S1 and Table S4). A random forest model for weight fraction bin was then built using function and property/use descriptors (5000 trees); the model error was estimated using 5-fold cross-validation and the model was tested using y-scrambling.

Predictive variable (descriptor) importance for both the function and weight fraction models was evaluated via a measure of the Gini importance [33], a mean (across all trees in the forest) of the decrease in the Gini impurity criterion (a measure of entropy) that results when a tree is split using a given descriptor as a classifier.

2.5. Application of the QSPR models for function and weight fraction to a case-study library of data-poor chemicals

The resulting QSPR classification models for function and weight fraction were applied to a library of chemicals having limited use and exposure data. This library included 10196 chemicals (including Tox21 chemicals) with either known use in PCPs but no weight fraction information (N = 538) or unknown specific use (N = 9658). The function and weight fraction models were applied to the first group; the function models were applied to the second. In our previous analysis of these thousands of chemicals with unknown use, the only available HT use heuristic shown to be correlated with exposures inferred from biomonitoring data was production volume [23].

The QSPR classification models were applied in a two-step manner to each data-poor chemical; each function model returned a probability (Pr) of the chemical having the function (equal to the fraction of the trees in the forest that returned a positive classification). Next, using the function associated with the highest probability, the weight fraction model was applied to predict a weight fraction bin (high/mid/low) for the chemical.

3. Results

3.1. Function-based empirical compositions for personal care product categories

The function and MSDS-based ingredient datasets were merged and summarized to develop empirical compositions in terms of function for the 66 PCP categories. Over 97% of the weight fraction observations could be matched to a function. The merged dataset comprised a total of 828 unique chemicals and 4115 individual PCPs, encompassing a total of 20975 records (weight-fraction/product category pairs) for use in generating empirical compositions. Harmonized function categories with the largest number of chemicals in the merged data were masking agents (N = 104), perfumes (N = 94), surfactants/cleansers/emulsifiers (N = 68), viscosity-controlling/emulsion stabilizers/binders (N = 60), emollients (N = 54), and colorants (N = 49). (Since the CosIng uses "perfume" as a function category label, we will later use "fragrances" to refer to products such as colognes, etc.)

The function-based empirical compositions for the 10 PCP categories with the largest number of unique products represented in the merged function-ingredient dataset are given in Table 1; functions appearing in at least 10% of the category formulations are reported. Compositions for all 66 categories (including additional percentiles) are given in SI Table S5. Weight fractions within a category that total more or less than 100% are due to variability across individual products or unreported ingredients on the MSDS. Median weight fractions across all functions and product categories are illustrated in Fig. 2. The highest median weight fractions across all product categories were found for solvents and skin conditioners, while in general the lowest weight fractions were found for colorants and preservatives.

The estimates of mean weight fraction for functions estimated in the current study compared favorably with available values derived by Dutch National Institute for Public Health and the Environment (RIVM) for use with the ConsExpo exposure model [15] for different PCP types [39] (SI Fig. S2); the RIVM values on the plot are means of reported values for products falling within the PCP categories used in this study. For several product categories, the RIVM values were higher than the current estimates. However, for 49 out of 67 product category-function pairs that could be compared, the RIVM value fell below the 95th percentile of the distributions we derived.

3.2. QSPR models for functions and weight fraction

The descriptors (properties and use) used in developing the classification models for function and weight fraction are listed in SI Table 3. The random forest model for predicting weight fraction passed the y-scrambling test (having an overall 5-fold cross validation error estimate of 16.7% compared to 45% obtained using scrambled weight fractions). The confusion matrix for the model is given in Table S6 of the SI; the largest potential for misclassification was high weight fractions being classified as medium. Function was the descriptor with the greatest importance in classifying weight fraction bin (SI Fig. S3), followed by molecular weight and vapor pressure. This result indicates that function is indeed relevant in predicting weight fraction in products. Production volume was the only use descriptor among the twelve highest-ranked predictors for weight fraction.

Descriptors (properties and use) were available for 2981 chemicals in the harmonized function dataset. A total of 26 functions had data for 10 or more chemicals; QSPR classification models were built for these functions. Of these 26 models, 22 were found to be acceptable using y-scrambling validation. These 22 models demonstrated good performance as measured by 5-fold cross-validation (Table S7). All 22 acceptable models had errors <19%; the best models were for propellants (1% error), colorants (6%) and oral-care/anti-plaque agents (4%). A model sensitivity and BA >70% was obtained for 19 and 21 functions, respectively; all function prediction models had specificity >82%. The predictive importance of the descriptors as measured by the Gini importance varied across function (illustrated in SI Fig. S4); this metric identifies the descriptors that had the most influence in predicting whether or not a chemical has a given function in the QSPR models. In general, the properties were much more important than the use descriptors, with the log of the octanol-water partition coefficent, boiling point, molecular weight, Henry's Law constant, and vapor pressure having relatively high Gini importance across many of the functions. The most influential use descriptor across functions was production volume (which, for example, was important for classifying solvents); the simple use descriptor for "colorant" was important for classifying the function "colorant" (indicating consistency across our use databases derived from different sources), as was the "personal care product use" descriptor in identifying preservatives, colorants, and perfumes.

The QSPR classification models were applied to a library of 10196 chemicals (including Tox21 library chemicals) having limited available exposure and use data. The heatmap in Fig. 3 illustrates predicted function probabilities (Pr) for chemicals with unknown specific use (N = 9658) and Pr and weight fraction bin predictions for those having known PCP use (N = 538). The darker bands indicate higher probability of a chemical having the function; the color of the bar on the far right side indicates weight fraction bin. Function was predicted with Pr > 0.9 for 197 of the PCP chemicals (37%), with chemicals most often predicted to be colorants, perfumes, preservatives, and emollients. For the chemicals with unknown use, function was predicted with Pr > 0.9 for 1360 chemicals (14%), with perfumes (640 chemicals), skin conditioners (299), colorants (35), and hair dyeing agents (73) making up a majority of the high probability results. Weight fractions were only predicted for chemicals with a maximum assigned Pr > 90%. Overall, based on their most likely predicted function, less than 1% of the weight fractions were predicted to be high weight fraction (30%-100% of total weight); 35% and 65% of the chemicals were respectively predicted to be in the medium and low weight fraction bins.

4. Discussion

The work presented in this paper is a step towards estimating in a HT fashion the functions and weight fractions of chemicals in one category of consumer products (PCPs). In the future this approach can be extended to additional types of products. The function and product categories developed here are not definitive; larger harmonized sets of function and consumer product categorizations would be useful in developing comprehensive databases of chemical-product-function sets for support of multiple modeling efforts. Barriers can exist to obtaining timely and abundant weight fraction data for exposure assessment (e.g. rapidly changing formulations and confidentiality concerns). We are currently working to expand and harmonize the CPCat [24] and CPCdb [26] databases to include additional chemical ingredient, function, and weight fraction information from a wider variety of sources, including additional manufacturer or retailer MSDS repositories, government[40] or industry[41] programs, and reported ingredient lists. We hope that providing this harmonized database to the public will encourage use by exposure assessors in both industry and government and promote further data sharing and transparency. These data can inform multiple tiers of exposure modeling – from targeted exposure assessments for single chemicals in single products, to HT (and high-uncertainty) approaches for chemical screening and prioritization. Targeted assessments are often more

Table 1
Empirical compositions (function-based weight fraction distributions) for the 10 personal care product categories having the largest number of unique products (N) represented in the merged function-ingredient data.

Category	Function	Percent of Formulations Containing Function	Weight Fraction			Number of Unique Chemicals Associated with Function	Most Common Chemical
			Mean	Median	SD		
BODY WASH (N = 150)	Solvents	40.67	0.468	0.200	0.422	6	AQUA (7732-18-5)
	Surfactants/Cleansers/Emulsifiers	68.67	0.075	0.050	0.070	19	SODIUM LAURETH SULFATE (9004-82-4)
	Viscosity-Controlling/Emulsion Stabilizers/Binding Agents	28.67	0.035	0.030	0.030	10	SODIUM CHLORIDE (7647-14-5)
	Buffering Agents	21.33	0.023	0.001	0.133	6	CITRIC ACID (77–92-9)
	Chelating Agents	22.67	0.021	0.010	0.051	4	TETRASODIUM EDTA (64-02-8)
	Masking Agents	33.33	0.018	0.010	0.036	18	TETRAMETHYL ACETYLOCTAHYDRONAPHTHALENES (54464-57-2)
	Perfumes	17.33	0.010	0.010	0.000	11	HEXYL CINNAMAL (101-86-0)
	Preservatives	24.67	0.005	0.001	0.012	14	METHYLPARABEN (99-76-3)
	Colorants	13.33	0.003	0.000	0.012	8	CI 19140 (1934-21-0)
FACE CREAM/MOISTURIZER (N = 154)	Solvents	47.40	0.191	0.100	0.254	6	GLYCERIN (56–81-5)
	Emollients	12.99	0.066	0.050	0.052	14	DIMETHICONE (9006-65-9)
	UV Absorbers/Filters	21.43	0.065	0.050	0.035	7	ETHYLHEXYL SALICYLATE (118-60-5)
	Antiseborroeic	32.47	0.047	0.050	0.010	1	NIACINAMIDE (98-92-0)
	Surfactants/Cleansers/Emulsifiers	13.64	0.033	0.015	0.034	8	TRIETHANOLAMINE (102-71-6)
	Viscosity-Controlling/Emulsion Stabilizers/Binding Agents	21.43	0.032	0.015	0.025	17	CETYL ALCOHOL (36653-82-4)
	Skin Conditioners	27.92	0.025	0.015	0.022	12	PENTYLENE GLYCOL (5343-92-0)
	Colorants	12.34	0.015	0.010	0.018	7	CI 77891 (13463-67-7)
	Preservatives	12.99	0.008	0.010	0.003	7	PHENOXYETHANOL (122-99-6)
FRAGRANCE (N = 150)	Ethyl Alcohol	47.33	0.883	0.980	0.188	1	ALCOHOL (64-17-5)
	Solvents	35.33	0.286	0.300	0.275	10	AQUA (7732-18-5)
	Surfactants/Cleansers/Emulsifiers	12.00	0.082	0.050	0.046	10	SODIUM LAURYL SULFATE (151-21-3)
	Masking Agents	47.33	0.025	0.010	0.034	55	TETRAMETHYL ACETYLOCTAHYDRONAPHTHALENES (54464-57-2)
	Skin Conditioners	14.67	0.025	0.010	0.028	3	ALPHA-ISOMETHYL IONONE (127-51-5)
	Perfumes	40.67	0.017	0.010	0.017	57	METHYLENEDIOXYPHENYL METHYLPROPANAL (1205-17-0)
	Tonics	13.33	0.007	0.010	0.004	3	GERANIOL (106-24-1)
HAIR COLOR (N = 108)	Buffering Agents	84.26	0.075	0.050	0.037	4	ETHANOLAMINE (141-43-5)
	Solvents	11.11	0.072	0.050	0.052	4	ISOPROPYL ALCOHOL (67-63-0)
	Viscosity-Controlling/Emulsion Stabilizers/Binding Agents	50.93	0.069	0.050	0.035	6	CETEARYL ALCOHOL (67762-27-0)
	Surfactants/Cleansers/Emulsifiers	83.33	0.062	0.050	0.029	11	OLETH-10 (9004-98-2)
	Antistatic Conditioners	41.67	0.052	0.050	0.009	3	SOYTRIMONIUM CHLORIDE (61790-41-8)
	Perfumes	35.19	0.040	0.010	0.052	6	TERPENES AND TERPENOIDS_ MIXED SOUR AND SWEET ORANGE OIL (68917-57-7)
	Hair Dyeing Agents	81.48	0.025	0.015	0.019	21	P-PHENYLENEDIAMINE (106-50-3)
	Masking Agents	17.59	0.013	0.010	0.012	3	D-LIMONENE (5989-27-5)
HAIR CONDITIONER (N = 141)	Solvents	24.11	0.373	0.020	0.463	5	AQUA (7732-18-5)
	Viscosity-Controlling/Emulsion Stabilizers/Binding Agents	19.15	0.047	0.050	0.026	4	CETYL ALCOHOL (36653-82-4)
	Antistatic Conditioners	75.89	0.045	0.050	0.013	11	STEARAMIDOPROPYL DIMETHYLAMINE (7651-02-7)

Table 1 (*Continued*)

Category	Function	Percent of Formulations Containing Function	Weight Fraction			Number of Unique Chemicals Associated with Function	Most Common Chemical
			Mean	Median	SD		
HAIR SPRAY (N = 128)	Ethyl Alcohol	88.28	0.547	0.550	0.218	1	ALCOHOL (64-17-5)
	Solvents	32.81	0.364	0.375	0.196	8	DIMETHYL ETHER (115-10-6)
	Propellants	32.81	0.277	0.250	0.154	4	HYDROFLUOROCARBON 152A (75-37-6)
	Film-Forming Agents	10.94	0.056	0.050	0.017	4	VA/CROTONATES/VINYL NEODECANOATE COPOLYMER (58748-38-2)
	Buffering Agents	20.31	0.014	0.010	0.012	2	AMINOMETHYL PROPANOL (124-68-5)
HAND/BODY LOTION (N = 143)	Solvents	72.73	0.443	0.200	0.413	3	GLYCERIN (56–81-5)
	Emollients	26.57	0.077	0.050	0.166	10	GLYCERYL STEARATE (31566-31-1)
	Skin Conditioners	28.67	0.059	0.050	0.133	8	PROPYLENE GLYCOL (57-55-6)
	Surfactants/Cleansers/Emulsifiers	37.76	0.055	0.020	0.162	17	TRIETHANOLAMINE (102-71-6)
	Viscosity-Controlling/Emulsion Stabilizers/Binding Agents	36.36	0.030	0.010	0.082	12	CETYL ALCOHOL (36653-82-4)
	Masking Agents	13.99	0.021	0.010	0.022	13	BENZOPHENONE (119-61-9)
	Chelating Agents	14.69	0.011	0.010	0.015	3	DISODIUM EDTA (139-33-3)
	Preservatives	36.36	0.009	0.010	0.011	13	METHYLPARABEN (99-76-3)
	Antioxidants	18.18	0.005	0.001	0.004	4	TOCOPHERYL ACETATE (7695-91-2)
	Colorants	13.29	0.002	0.001	0.004	8	CI 77891 (13463-67-7)
NAIL POLISH (N = 117)	Solvents	78.63	0.310	0.300	0.223	16	ETHYL ACETATE (141-78-6)
	Ethyl Alcohol	17.95	0.287	0.300	0.087	1	ALCOHOL (64-17-5)
	Film-Forming Agents	80.34	0.169	0.100	0.204	14	NITROCELLULOSE (9004-70-0)
	Plasticizers	18.80	0.048	0.045	0.027	3	TRIMETHYL PENTANYL DIISOBUTYRATE (6846-50-0)
	Viscosity-Controlling/Emulsion Stabilizers/Binding Agents	34.19	0.045	0.010	0.044	10	PVP (9003-39-8)
	Masking Agents	29.06	0.034	0.020	0.030	5	CAMPHOR (76-22-2)
	Surfactants/Cleansers/Emulsifiers	16.24	0.010	0.010	0.007	2	TRIETHANOLAMINE (102-71-6)
	UV Absorbers/Filters	13.68	0.010	0.005	0.015	3	BENZOPHENONE-1 (131-56-6)
	Emollients	11.97	0.009	0.010	0.005	3	DIMETHICONE (9006-65-9)
	Colorants	44.44	0.006	0.000	0.019	26	CI 77891 (13463-67-7)
	Preservatives	20.51	0.006	0.010	0.004	7	DMDM HYDANTOIN (6440-58-0)
SHAMPOO (N = 242)	Solvents	15.70	0.523	0.630	0.381	7	AQUA (7732-18-5)
	Surfactants/Cleansers/Emulsifiers	86.36	0.085	0.100	0.028	23	SODIUM LAURYL SULFATE (151-21-3)
	Viscosity-Controlling/Emulsion Stabilizers/Binding Agents	23.14	0.039	0.050	0.022	6	COCAMIDE MEA (68140-00-1)
	Buffering Agents	18.18	0.029	0.030	0.021	5	CITRIC ACID (77–92-9)
SHAVING CREAM, GEL (N = 102)	Surfactants/Cleansers/Emulsifiers	29.41	0.077	0.070	0.022	3	TRIETHANOLAMINE (102-71-6)
	Solvents	98.04	0.076	0.050	0.152	4	ISOPENTANE (78-78-4)
	Propellants	33.33	0.022	0.010	0.019	2	ISOBUTANE (75-28-5)
	Skin Conditioners	16.67	0.016	0.005	0.025	2	PROPYLENE GLYCOL (57-55-6)
	Viscosity-Controlling/Emulsion Stabilizers/Binding Agents	10.78	0.011	0.010	0.011	4	PTFE (9002-84-0)
	Perfumes	24.51	0.010	0.010	0.000	2	4-*tert*-BUTYLCYCLOHEXYL ACETATE (32210-23-4)

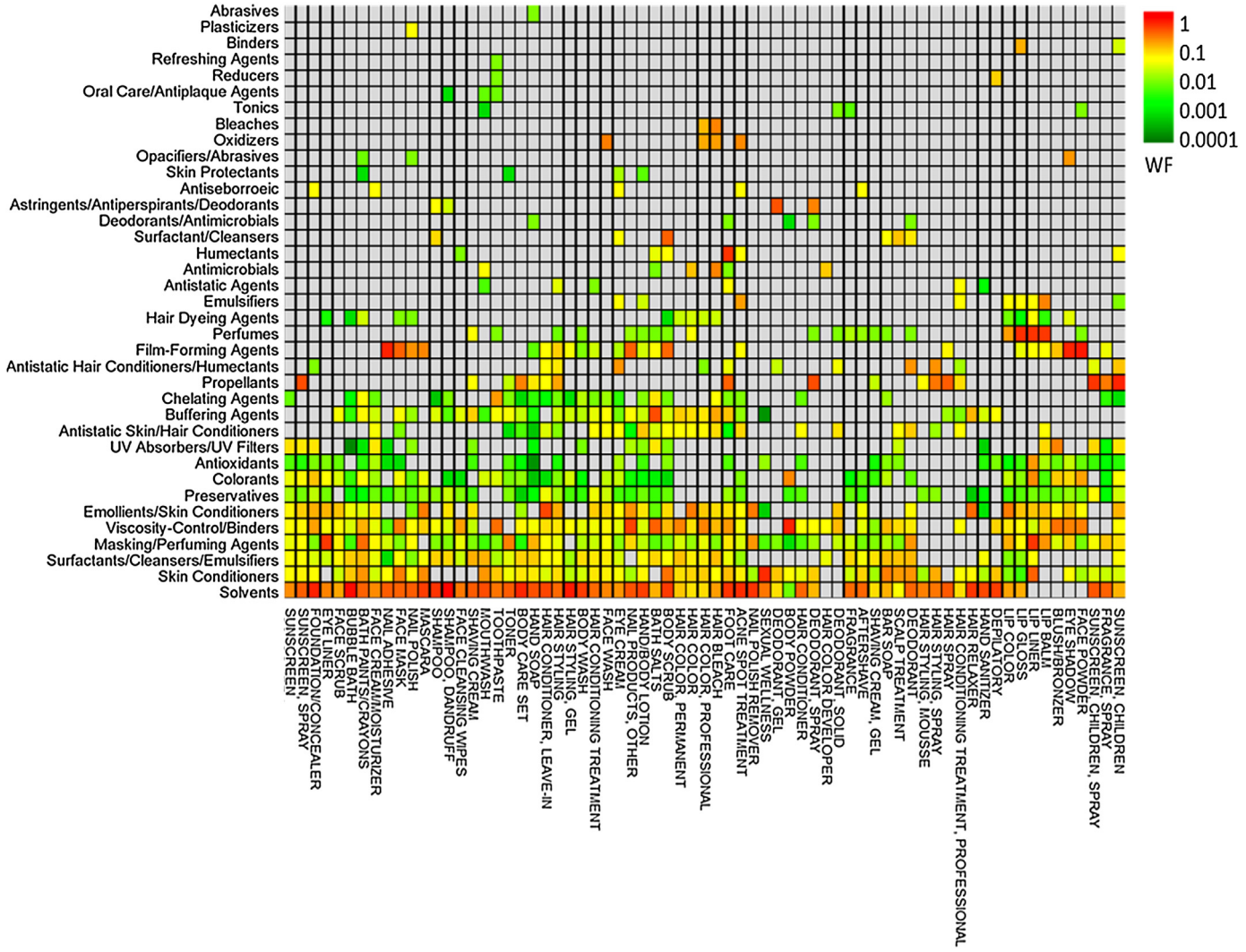

Fig. 2. Median product weight fractions (WF) associated with 35 harmonized chemical functions in 66 personal care product categories. Heatmap is clustered on both axes to group products/functions that are similar in terms of WF.

data-rich: specific product concentration data may be available, and time and resources allow for a greater examination of the processes that ultimately lead to human exposures. However, HT low-tier models that are most fit for prioritization must rely heavily on data that is readily available for thousands of chemicals. For these efforts, models such as those developed here are appropriate for filling gaps in available information.

4.1. Function-based empirical compositions for personal care products

Quantitative estimates of chemical consumer product weight fractions for chemicals are used in many chemical assessment programs. However, while some generic or "framework" compositions (i.e. formulations) are available, specific information about the variability in concentrations over many formulations (for example for use in Monte Carlo-based assessments) has been scarce. The work described here builds on and augments currently available information by developing generic compositions from data on thousands of real PCPs in commerce, thereby adding variability to current generic product formulations. These compositions will be useful for predicting exposures for known chemicals in PCPs for which we have no quantitative weight fraction information and for new chemicals with known uses. Further improving the characterization of variability in chemical concentrations across products in a single category and across multiple product categories can improve estimates of aggregate and cumulative exposures.

While we were able to generate useful compositions for many product categories, this exercise did identify some potential limitations in ingredient databases derived from MSDS sources. MSDS reporting rules are such that ingredient information is often incomplete; chemicals with weight fractions below a given threshold may be excluded (in general 1%, but 0.1% for identified carcinogens) [42]. Data gaps in some of the empirical compositions were identified. The most obvious example of this was the PCP category "Fragrances" (e.g., colognes). A majority of the 150 fragrance products in the weight fraction dataset (83.86%) contained a solvent (including ethyl alcohol), while only 40% reported containing a chemical identified having the function "perfume." It is likely that this data gap results from the failure to report specific chemicals on the MSDS due to the low concentrations of the chemicals, lack of information about fragrance mixtures purchased from third parties, or the absence of toxicity data. Formulations are also confidential, and no U.S. regulation requires the disclosure of any ingredient in a fragrance mixture, or of all ingredients in consumer products [43]. This kind of underreporting likely occurs for other functions as well. Future data collection efforts should seek to identify and analyze additional information sources beyond MSDS sheets, including

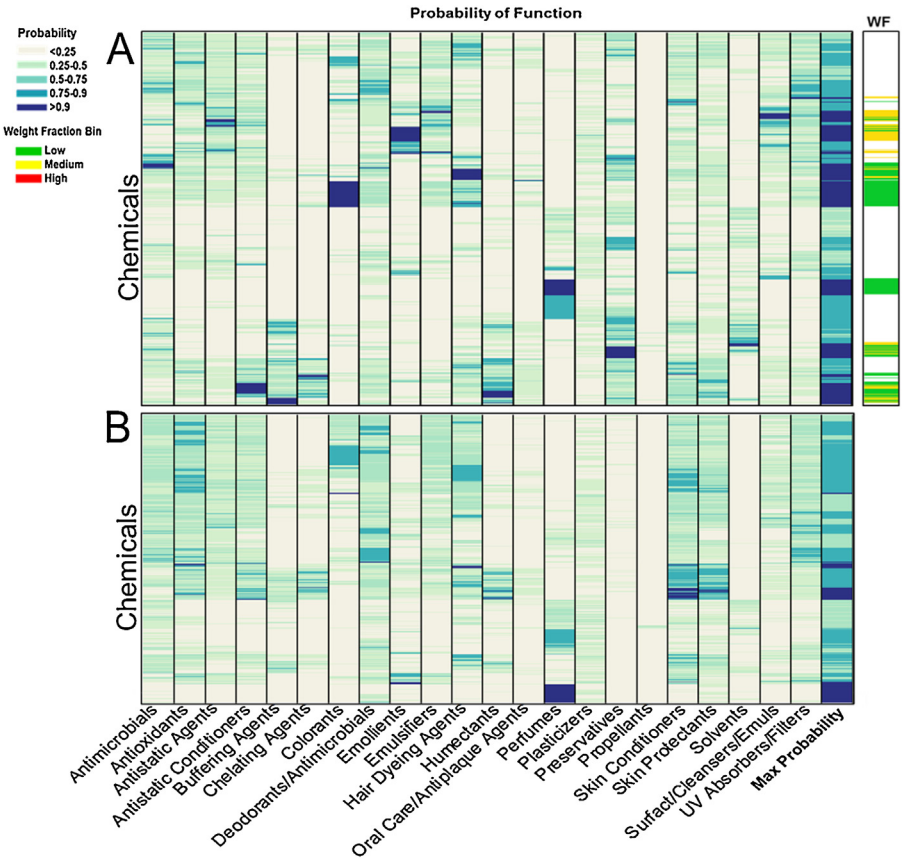

Fig. 3. Function and weight fraction bin predictions for chemicals having identified personal care product use (A; N = 538) and function predictions for chemicals with unknown use (B; N = 9658). Chemical observations are on the vertical axis; the two datasets are scaled differently on this axis for size. Chemicals having a high probability are shown in dark blue. Weight fraction predictions (indicated by the colorbar to the right of the main figure) are reported when a function could be predicted with >90% probability. Only functions are predicted for the unknown use chemicals since they may or may not be present in consumer products.

publicly-available product ingredient lists from manufacturers or retailers. Such sources have recently been used to identify chemical mixtures in consumer products [44].

4.2. QSPR prediction of functions and consumer product weight fraction

The QSPR chemical function and weight fraction models derived here can be used to parameterize existing mechanistic exposure models for PCPs (such as SHEDS-HT [13] or others [45,46]). In our MSDS ingredient database, over 26000 additional ingredient observations (CASRN linked to PCP category) are available that are missing quantitative weight fractions. Predicted functions for these chemicals can be used with the empirical compositions to directly fill in these data for use in running SHEDS-HT (effectively doubling the size of our product-weight fraction dataset). In addition, over 37% of the PCP chemicals in the data-poor library could be assigned functions with a probability >90%; for these chemicals, this function classification will enable prediction of the exposures associated with each relevant PCP category using the empirical compositions. If one assumes presence of a chemical in all PCP types (since these chemicals are not linked to a specific PCP category), conservative estimate of aggregate exposures can be made using SHEDS. Within SHEDS-HT, over 200 consumer product categories are directly linked to product use patterns (e.g. magnitude and frequency of use), exposure scenarios (e.g. direct dermal application), and exposure route (e.g. dermal absorption, inhalation, or hand-to-mouth transfer); when parameterized with quantitative weight fractions, these algorithms return estimates of total chemical intake in mg/kg/day. These exposure predictions (and the results of other models) can be combined with oral equivalent doses estimated from HT bioactivity assays to produce metrics for prioritizing chemicals on the basis of risk [47].

The QSPR function models can also be used to generate refined chemical use heuristics for predictive exposure modeling. For the library of chemicals with unknown uses, we are now able to assign functional use descriptors (e.g. “dye”, “colorant”, “perfume”) for over 600 chemicals for which we had no available use data. Although the chemicals could be used in the predicted functional role in other sectors besides consumer products, the predicted function adds additional information about chemical use that was previously unavailable. These new descriptors can be used in the development and application of empirical exposure heuristic models such as those based on regression against exposures inferred from biomarker data [23]. In addition, these descriptors will be useful in analyzing chemicals identified in HT non-targeted or suspect screening of environmental media [48] or consumer product formulations, such as efforts ongoing under ExpoCast. Such analyses could provide evidence of other exposure sources (such as contaminants).

The classifier models presented here are not meant to be definitive but rather demonstrate the potential utility of such a modeling framework. These models can be improved by the inclusion of additional chemicals with identified functions into the training sets, and identification of additional chemical-specific descriptors (for improved discrimination among functions). For example, only 14% of the chemicals with unknown use could be assigned function with a probability of >90%; the creation of additional models for other

functions outside of those used in consumer products will likely increase this percentage since many of these chemicals are likely used in other sectors. However, the current analysis demonstrates a promising path forward for identifying 1) the function of arbitrary chemicals and 2) weight fractions for data-poor chemicals known to be in consumer products. The analysis here was limited to PCPs since a large source of function data for these types of products (the CosIng database) was available. However, the prediction methodology can easily be expanded to more chemical functions that cover additional consumer product, article, or industrial categories. Other information on functional use for chemicals is currently being developed by government [40] and industry [49,50] programs and additional classification models can consider sector of use in addition to function (e.g., to differentiate between solvents in cleaning products versus solvents in PCPs). New classification models can also include additional descriptors available for thousands of chemicals, including structural information such as the ToxPrint chemotypes, a public set of over 700 structural fragments developed for data mining and modeling [51]. These new function and descriptor data sources can improve the accuracy of the current function and weight fraction predictor models. Refined QSPR function classification models can inform function-based chemical alternatives assessment [52] and molecular repurposing by identifying potential alternatives from large chemical libraries on the basis of predicted function.

4.3. Functional role and exposure potential of chemicals

The Organisation for Economic Co-operation and Development's (OECD) Guidance on the Grouping of Chemicals [53] emphasizes the utility of chemical grouping with respect to adverse outcome pathways or toxicological endpoints for the purpose of filling data gaps. Grouping by function could be useful for analogous read-across in terms of exposure potential, as functional role can determine the types of consumer products or articles containing chemicals and the concentrations in which they are present (e.g., flame retardants are primarily present in furnishings and clothing, while fragrances occur across many categories of PCPs and cleaning products). This read-across could be applied to exposure factors, exposure measurement (monitoring) data, or exposure model predictions. The ability to perform such read-across will be required for the development of robust assessments that consider aggregate (multi-pathway, multi-scenario, multi-product) exposures for single chemicals and ultimately cumulative assessments that consider groups of chemicals with similar hazard endpoints.

5. Conclusions

Qualitative and quantitative consumer product chemical ingredient information is critical input to the exposure component of HT risk-based analysis of chemicals. Such ingredient information is relevant to multiple tiers of assessment, including screening of large numbers of chemicals on the basis of exposure potential, identification of plausible exposure pathways for families of chemicals, and development of weight fractions ranges for use in detailed exposure assessments of single products, product families, or substances. The methods presented here make use of available information from thousands of real products and chemicals in commerce to build PCP ingredient profiles and predictive chemical classification models on the basis of the functional roles that chemicals perform. These methods comprise a straightforward and standardized approach for filling gaps in existing consumer product ingredient data for use in HT chemical prioritization on the basis of exposure.

6. Transparency document

The Transparency document associated with this article can be found in the online version.

7. Conflict of interest

The authors declare no conflict of interests.

Acknowledgements

The work reported here was funded by the US Environmental Protection Agency, in part under contract EP-C-14-001 to ICF International, Inc. Its contents are solely the authors' responsibility and do not necessarily represent official views of the Agency. The paper has been subjected to the Agency's review process and approved for publication. The Oak Ridge Institute for Science and Education provided funding for K. Phillips. The authors would like to thank Drs. Paul Price and Brandall Ingle for thoughtful and helpful review of the manuscript, and the EPA Chemical Safety for Sustainability program for support of this research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.toxrep.2016.08.011.

References

[1] E.A. Cohen-Hubal, A. Richard, L. Aylward, S. Edwards, J. Gallagher, M.-R. Goldsmith, S. Isukapalli, R. Tornero-Velez, E. Weber, R. Kavlock, Advancing exposure characterization for chemical evaluation and risk assessment, J. Toxicol. Environ. Health 13 (2010) 299–313.
[2] Exposure Science in the 21st Century: A Vision and a Strategy, National Research Council of the National Academies. National Academies Press, 2012.
[3] M.A. Jayjock, C.F. Chaisson, C.A. Franklin, S. Arnold, P.S. Price, Using publicly available information to create exposure and risk-based ranking of chemicals used in the workplace and consumer products, J. Expo. Sci. Environ. Epidemiol. 19 (2009) 515–524.
[4] R.A. Rudel, D.E. Camann, J.D. Spengler, L.R. Korn, J.G. Brody, Phthalates, alkylphenols, pesticides, polybrominated diphenyl ethers, and other endocrine-disrupting compounds in indoor air and dust, Environ. Sci. Technol. 37 (2003) 4543–4553.
[5] R.E. Dodson, M. Nishioka, L.J. Standley, L.J. Perovich, J.G. Brody, R.A. Rudel, Endocrine disruptors and asthma-associated chemicals in consumer products, Environ. Health Perspect. 120 (2012) 935–943.
[6] T. Schettler, Human exposure to phthalates via consumer products, Int. J. Androl. 29 (2006) 134–139.
[7] C.J. Weschler, W.W. Nazaroff, Semivolatile organic compounds in indoor environments, Environ. Sci. Technol. 42 (2008) 9018–9040.
[8] R.A. Rudel, L.J. Perovich, Endocrine disrupting chemicals in indoor and outdoor air, Atmos. Environ. 43 (2009) 170–181.
[9] G.A. Glegg, J.P. Richards, Chemicals in household products: problems with solutions, Environ. Manag. 40 (2007) 889–901.
[10] J.F. Wambaugh, R.W. Setzer, D.M. Reif, S. Gangwal, J. Mitchell-Blackwood, J.A. Arnot, O. Joliet, A. Frame, J. Rabinowitz, T.B. Knudsen, R.S. Judson, P. Egeghy, D. Vallero, E.A. Cohen Hubal, High-throughput models for exposure-based chemical prioritization in the ExpoCast project, Environ. Sci. Technol. 47 (2013) 8479–8488.
[11] L.A. Wallace, Comparison of risks from outdoor and indoor exposure to toxic chemicals, Environ. Health Perspect. 95 (1991) 7–13.
[12] W.R. Ott, Total human exposure: basic concepts, EPA field studies, and future research needs, J. Air Waste Manag. Assoc. 40 (1990) 966–975.
[13] K.K. Isaacs, W.G. Glen, P. Egeghy, M.R. Goldsmith, L. Smith, D. Vallero, R. Brooks, C.M. Grulke, H. Özkaynak, SHEDS-HT: an integrated probabilistic exposure model for prioritizing exposures to chemicals with near-field and dietary sources, Environ. Sci. Technol. 48 (2014) 12750–12759.
[14] C. Delmaar, B. Bokkers, W. ter Burg, G. Schuur, Validation of an aggregate exposure model for substances in consumer products: a case study of diethyl phthalate in personal care products, J. Expo. Sci. Environ. Epidemiol. 25 (2015) 317–323.
[15] J. Delmaar, M. Park, J. van Engelen, ConsExpo 4.0 Consumer Exposure and Uptake Models Program Manual, RIVM Report 320104004/2005, National Institute for Public Health and the Evironment, Bilthoven, The Netherlands, 2005.

[16] Consumer Exposure Model, United States Environmental Protection Agency, 2015 http://www.epa.gov/sites/production/files/2015-09/documents/cem_user_guide_beta_test.pdf.
[17] Description of Existing Models and Tools Used for Exposure Assessment, Organisation for Economic Co-operation and Development ENV/JM/MONO(2012)37, 2012 http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2012)37&doclanguage=en.
[18] Consumer Product Ingredient Safety: Exposure and Risk Screening Methods for Consumer Product Ingredients, 2nd edition, American Cleaning Institute, 2010 http://www.aciscience.org/docs/Consumer_Product_Ingredient_Safety_v2.0.pdf.
[19] P.P. Egeghy, R. Judson, S. Gangwal, S. Mosher, D. Smith, J. Vail, E.A. Cohen Hubal, The exposure data landscape for manufactured chemicals, Sci. Total Environ. 414 (2012) 159–166.
[20] Chemical Requirements for Consumer Products: Proposals for Regulatory Measures to Improve Chemical Safety for Consumers, Austrian Federal Ministry of Labour, Social Affairs and Consumer Protection, 2010 http://www.verbraucherrat.at/content/01-news/06-archiv-2011-2012/01-forschungsarbeit-zu-chemikalien-in-verbraucherprodukten/chemicalsproducts1.pdf.
[21] D.J. Dix, K.A. Houck, M.T. Martin, A.M. Richard, R.W. Setzer, R.J. Kavlock, The ToxCast program for prioritizing toxicity testing of environmental chemicals, Toxicol. Sci. 95 (2007) 5–12.
[22] R.R. Tice, C.P. Austin, R.J. Kavlock, J.R. Bucher, Improving the human hazard characterization of chemicals: a Tox21 update, Environ. Health Perspect. 121 (7) (2013) 756–765.
[23] J.F. Wambaugh, A. Wang, K.L. Dionisio, A. Frame, P. Egeghy, R. Judson, R.W. Setzer, High throughput heuristics for prioritizing human exposure to environmental chemicals, Environ. Sci. Technol. 48 (21) (2014) 12760–12767.
[24] K. Dionisio, A.F. Frame, M.-R. Goldsmith, J.F. Wambaugh, A. Liddell, T. Cathey, Exploring consumer exposure pathways and patterns of use for chemicals in the environment, Toxicol. Rep. 2 (2015) 228–237.
[25] The Chemical/Product Categories Database (CPCat), http://actor.epa.gov/cpcat/faces/home.xhtml. (accessed 14.06.16).
[26] M.R. Goldsmith, C.M. Grulke, R.D. Brooks, T.R. Transue, Y.M. Tan, A. Frame, P.P. Egeghy, R. Edwards, D.T. Chang, R. Tornero-Velez, K. Isaacs, A. Wang, J. Johnson, K. Holm, M. Reich, J. Mitchell, D.A. Vallero, L. Phillips, M. Phillips, J.F. Wambaugh, R.S. Judson, T.J. Buckley, C.C. Dary, Development of a consumer product ingredient database for chemical exposure screening and prioritization, Food Chem. Toxicol. 65 (2014) 269–279.
[27] G. Chevillotte, A.S. Ficheux, T. Morisset, A.C. Roudot, Exposure method development for risk assessment to cosmetic products using a standard composition, Food Chem. Toxicol. 68 (2014) 108–116.
[28] CosIng: Cosmetic Ingredients and Substances, European Commission, http://ec.europa.eu/growth/tools-databases/cosing/. (accessed 14.06.16).
[29] Official Journal of the European Union. Regulation (EC) No. 1223/2009 of the European Parliament and of the Council of 30 November 2009 on Cosmetic Products, http://eur-ex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:342:0059:0209:en:PDF.32.
[30] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley, New York, NY, 2005.
[31] K. Roy, S. Kar, R. Das, Statistical methods in QSAR/QSPR, in: A Primer on QSAR/QSPR Modeling: Fundamental Concepts, Springer International Publishing, Cham, 2015, pp. 37–59.
[32] USEPA Estimation Programs Interface Suite™ for Microsoft® Windows, V 4.11 http://www.epa.gov/oppt/exposure/pubs/episuite.htm.
[33] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[34] The R Project for Statistical Computing (accessed 14.06.16) http://www.r-project.org.
[35] R. Package Random Forest., 2016 (accessed 14.06.16) http://cran.r-project.org/web/packages/randomForest/index.html.
[36] C. Chen, A. Liaw, L. Breiman, Using Random Forests to Learn Imbalanced Data, University of California Berkeley, Department of Statistics, 2004, Report 666 http://statistics.berkeley.edu/tech-reports/666.
[37] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, New York, 2013.
[38] A. Tropsha, P. Gramatica, V. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003), 1611–0218.
[39] H. Bremmer, L. Prud'homme de Lodder, Van Engelen, J. Cosmetics Fact Sheet. To Assess the Risks for the Consumer. Updated Version for ConsExpo 4. RIVM Report 320104001/2006, National Institute for Public Health and the Environment, Bilthoven, The Netherlands, 2006.
[40] U.S. EPA Design for the Environment Website (accessed 14.06.16) http://www.epa.gov/dfe.
[41] Cosmetic Ingredient Review Database (accessed 10.09.16) http://www.cir-safety.org/ingredients.
[42] Hazard Communication. 29 C.F.R. 1910.1200. United States Code of Federal Regulations. http://www.ecfr.gov/cgi-bin/text-idx?SID=860a2ccbbd906c0480a386908a241315&mc=true&node=se29.6.1910_11200&rgn=div8.
[43] A.C. Steinemann, Fragranced consumer products and undisclosed ingredients, Environ. Impact Assess. Rev. 29 (1) (2009) 32–38.
[44] H.A. Gabb, C. Blake, An informatics approach to evaluating combined chemical exposures from consumer products: a case study of asthma-associated and potential endocrine disruptors, Environ. Health Perspect. (2016), E-pub ahead of print.
[45] Exposure and Fate Assessment Screening Tool Version, 2014 http://www.epa.gov/oppt/exposure/pubs/efast.htm.
[46] X. Zhang, J.A. Arnot, F. Wania, Model for screening-level assessment of near-field human exposure to neutral organic chemicals released indoors, Environ. Sci. Technol. 48 (2014) 12312–12319.
[47] U.S. EPA, Integrated Bioactivity and Exposure Ranking: A Computational Approach for the Prioritization and Screening of Chemicals in the Endocrine Disruptor Screening Program, 2014 https://www.regulations.gov/#!documentDetail,D=EPA-HQ-OPP-2014-0614-0003.
[48] J.E. Rager, M.J. Strynar, S. Liang, R.L. McMahen, A.M. Richard, C.M. Grulke, J.F. Wambaugh, K.K. Isaacs, R. Judson, A.J. Williams, J.R. Sobus, Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring, Environ. Int. 88 (2016) 269–280.
[49] Consumer Specialty Products Association Consumer Products Ingredient Dictionary, 2016 http://www.cspa.org/product/consumer-products-ingredient-dictionary/.
[50] CleanGredients (accessed 10.09.2016) http://www.cleangredients.org/.
[51] C. Yang, A. Tarkhov, J. Marusczyk, B. Bienfait, J. Gasteiger, T. Kleinoeder, T. Magdziarz, O. Sacher, C.H. Schwab, J. Schwoebel, L. Terfloth, K. Arvidson, A. Richard, A. Worth, J. Rathman, New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling, J. Chem. Inf. Model. 55 (3) (2015) 510–528.
[52] J.A. Tickner, J.N. Schifano, A. Blake, C. Rudisill, M.J. Mulvihill, Advancing safer alternatives through functional substitution, Environ. Sci. Technol. 49 (2015) 742–749.
[53] Organisation for Economic Co-operation and Development, Guidance on Grouping of Chemicals, second edition, OECD, Paris, 2014, Series on Testing & Assessment No. 194. ENV/JM/MONO(2014)4.