AN INFORMATICS APPROACH TO PRIORITIZING RISK ASSESSMENT FOR CHEMICALS AND CHEMICAL COMBINATIONS BASED ON NEAR-FIELD EXPOSURE FROM CONSUMER PRODUCTS

BY

HENRY A GABB

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Associate Professor Catherine Blake, Chair and Director of Research
Professor Allen Renear
Professor Jodi Flaws
Dr. Ian Brooks
Associate Professor Nathaniel Osgood, University of Saskatchewan

Abstract

Over 80,000 chemicals are registered under the U.S. Toxic Substances Control Act of 1976, but only a few hundred have been screened for human toxicity. Not even those used in everyday consumer products, and known to have widespread exposure in the general population, have been screened. Toxicity screening is time-consuming, expensive, and complex because simultaneous or sequential exposure to multiple environmental stressors can affect chemical toxicity. Cumulative risk assessments consider multiple stressors but it is impractical to test every chemical combination and environmental stressor to which people are exposed. The goal of this research is to prioritize the chemical ingredients in consumer products and their most prevalent combinations for risk assessment based on likely exposure and retention.

This work is motivated by two concerns. The first, as noted above, is the vast number of environmental chemicals with unknown toxicity. Our body burden (or chemical load) is much greater today than a century ago. The second motivating concern is the mounting evidence that many of these chemicals are potentially harmful. This makes us the unwitting participants in a vast, uncontrolled biochemistry experiment.

An informatics approach is developed here that uses publicly available data to estimate chemical exposure from everyday consumer products, which account for a significant proportion of overall chemical load. Several barriers have to be overcome in order for this approach to be effective. First, a structured database of consumer products has to be created. Even though such data is largely public, it is not readily available or easily accessible. The requisite consumer product information is retrieved from online retailers. The resulting database contains brand, name, ingredients, and category for tens of thousands of unique products. Second, chemical nomenclature is often ambiguous. Synonymy (i.e., different names for the same chemical) and homonymy (i.e., the same name for different chemicals) are rampant. The PubChem Compound database, and to a lesser extent the Universal Medical Language System, are used to map chemicals to unique identifiers. Third, lists of toxicologically interesting chemicals have to be compiled. Fortunately, several authoritative bodies (e.g., the U.S. Environmental Protection Agency) publish lists of suspected harmful chemicals to be prioritized for risk assessment. Fourth, tabulating the mere presence of potentially harmful

chemicals and their co-occurrence within consumer product formulations is not as interesting as quantifying likely exposure based on consumer usage patterns and product usage modes, so product usage patterns from actual consumers are required. A suitable dataset is obtained from the Kantar Worldpanel, a market analysis firm that tracks consumer behavior. Finally, a computationally feasible probabilistic approach has to be developed to estimate likely exposure and retention for individual chemicals and their combinations. The former is defined here as the presence of a chemical in a product used by a consumer. The latter is exposure combined with the relative likelihood that the chemical will be absorbed by the consumer based on a product's usage mode (e.g., whether the product is rinsed off or left on after use).

The results of four separate analyses are presented here to show the efficacy of the informatics approach. The first is a proof-of-concept demonstrating that the first two barriers, creating the consumer product database and dealing with chemical synonymy and homonymy, can be overcome and that the resulting system can measure the per-product prevalence of a small set of target chemicals (55 asthma-associated and endocrine disrupting compounds) and their combinations. A database of 38,975 distinct consumer products and 32,231 distinct ingredient names was created by scraping Drugstore.com, an online retailer. Nearly one-third of the products (11,688 products, 30%) contained ≥1 target chemical and 5,229 products (13%) contained >1. Of the 55 target chemicals, 31 (56%) appear in ≥1 product and 19 (35%) appear under more than one name. The most frequent 3-way chemical combination (2-phenoxyethanol, methyl paraben, and ethyl paraben) appears in 1,059 products.

The second analysis demonstrates that the informatics approach can scale to several thousand target chemicals (11,964 environmental chemicals compiled from five authoritative lists). It repeats the proof-of-concept using a larger product sample (55,209 consumer products). In the third analysis, product usage patterns and usage modes are incorporated. This analysis yields unbiased, rational prioritizations of potentially hazardous chemicals and chemical combinations based on their prevalence within a subset of the product sample (29,814 personal care products), combined exposure from multiple products based on actual consumer behavior, and likely chemical retention based on product usage modes. High-ranking chemicals, and combinations thereof, include glycerol; octamethyltrisiloxane; citric acid;

titanium dioxide; 1,2-propanediol; octadecan-1-ol; saccharin; hexitol; limonene; linalool; vitamin e; and 2-phenoxyethanol. The fourth analysis is the same as the third except that each authoritative list is prioritized individually for side-by-side comparison.

The informatics approach is a viable and rationale way to prioritize chemicals and chemical combinations for risk assessment based on near-field exposure and retention. Compared to spectrographic approaches to chemical detection, the informatics approach has the advantage of a larger product sample, so it often detects chemicals that are missed during spectrographic analysis. However, the informatics approach is limited to the chemicals that are actually listed on product labels. Manufacturers are not required to specify the chemicals in fragrance or flavor mixtures, so the presence of some chemicals may be underestimated. Likewise, chemicals that are not part of the product formulation (e.g., chemicals leached from packaging, degradation byproducts) cannot be detected. Therefore, spectrographic and informatics approaches are complementary.

Acknowledgements

This work would not have been possible without the support of my family, friends, and teachers. Special thanks go to my wife, Mary, who has been through this process once before. I know it didn't get any easier. Thank you, Grace, my favorite daughter, for your patience when I was too busy to play.

I will miss my classmates in the 2012 doctoral cohort – some of the smartest people I have ever met. Our discussions about library and information science in and out of class eased my reentry into scholarly work. I hope our paths cross in future work. I will also miss my lab mates; particularly, Ana Lucic, linguist extraordinaire, and Jenna Kim, who meticulously checked my code and workflows. I wish you all the best of luck in your research careers.

I would like to thank the faculty and staff of the School of Information Sciences, especially my instructors. I learned something valuable in every course. In particular, I would like to thank Dr. Alistair Black, who taught my favorite class (History and Foundations of Library and Information Science) and coauthored my first history article. If we ever get some spare time, let's study how Britain's experience during "The Blitz" informed American plans to protect information and cultural assets during the Cold War. I'll need your skill at archival research. I would also like to thank my late committee member, Dr. Les Gasser, for his discussions of Popper, Kuhn, the flaws of logical positivism, and most importantly, his insistence upon explicit data validation during my research. I will miss your insight.

I would like to thank my committee for guiding this research. Your ideas and suggestions were invaluable. I would also like to thank Drs. Susan Schantz and Andrea Aguiar of the Children's Environmental Health Research Center for their helpful discussions in the early phases of this research. Finally, I would like to extend my gratitude to my advisor, Dr. Cathy Blake, who taught me everything I know about text mining and natural language processing. Perhaps I'll become proficient at Java programming someday, but don't count on it.

Table of Contents

List of Figures viii
List of Tables ix
Chapter 1: Introduction and Motivation 1
Chapter 2: Related Work 9
2.1 Generating Consumer Product Databases 9
2.2 Prioritizing Chemicals for Risk Assessment 10
Chapter 3: Data Collection 14
3.1 Selecting and Preprocessing the Target Chemicals 14
3.2 Selecting the Chemical Dictionary 22
3.3 Creating a Database of Consumer Products 26
3.3.1 Verifying that the Data Owner Allows Scraping 26
3.3.2 Scraping the Online Retailer 27
3.3.3 Extracting the Required Information from the Raw HTML 29
3.4 Incorporating Consumer Product Usage Patterns and Usage Modes 39
Chapter 4: Data Cleaning 42
4.1 Cleaning the Consumer Product Database 42
4.1.1 Removing Duplicate Products from the Database 42
4.1.2 Assigning Product Categories 45
4.1.3 Tabulating the Product Sets 49
4.2 Cleaning and Using the Chemical Dictionary 52
4.2.1 Mapping the Target Chemicals to Unique Identifiers 52
4.2.2 Mapping Product Ingredients to Unique Identifiers 58
4.2.3 Resolving Chemical Synonymy 65
4.2.4 Accounting for Chemical Homonymy 68
Chapter 5: Chemical Exposure and Retention Factors 72
Chapter 6: Prevalent Target Chemicals in Consumer Products 76
6.1 Ranking Prevalent Chemicals and Chemical Combinations in Consumer Products 76
6.2 Ranking Prevalent Chemicals and Chemical Combinations among Consumers 86
6.2.1 Ranking Chemicals and Chemical Combinations by EF and RF 86
6.2.2 Qualitatively and Quantitatively Comparing the Ranked Lists 93

6.3 Rankings for Each Authoritative List 96
6.3.1 Ranking the Tox21 10K Library 96
6.3.2 Ranking the Hazardous Substances Data Bank 100
6.3.3 Ranking the California Chemicals of Concern 104
6.3.4 Ranking the Endocrine Disrupting Compounds Data Bank 107
6.3.5 Ranking the Compounds from Dodson et al. (2012) 111
Chapter 7: Discussion 116
7.1 Assessing the Authoritative Lists of Potentially Harmful Chemicals 116
7.2 Limitations of the Informatics Approach 119
7.3 Evaluating the Ranked Lists without *A Priori* Knowledge of Biological Activity 120
Chapter 8: Conclusions and Future Work 124
8.1 Effectiveness of Informatics Approaches to Near-Field Chemical Exposure 124
8.2 Future Work 127
8.2.1 Incorporating HTS Data from ToxCast 127
8.2.2 Using Chemical Absorption Models When Computing RF 130
8.2.3 Taking Advantage of Unused Product Data 130
8.2.4 Reengineering the Informatics Workflow 132
References 138
Appendix: Supplemental Material 151

List of Figures

Figure 1 Swiss cheese model of adverse effects 3
Figure 2 Research goal 7
Figure 3 Preprocessing the authoritative lists of target chemicals 20
Figure 4 Web scraping process 28
Figure 5 Example product webpage from Drugstore.com 30
Figure 6 Example of raw HTML for Drugstore.com product webpages 31
Figure 7 Extracting and cleaning data from the raw consumer product HTML files 36
Figure 8 Structure of the Kantar Worldpanel dataset used in this research (not actual data) 40
Figure 9 Algorithm used to find duplicate products 44
Figure 10 Assigning categories to the consumer products 47
Figure 11 Mapping the target chemicals to PubChem CIDs 54
Figure 12 Overlap among the authoritative lists of potentially harmful chemicals 58
Figure 13 Generating the matrix of target chemical proportions by product category 60
Figure 14 Example of homonymy in chemical naming 69
Figure 15 Probabilistic algorithm to prioritize the target chemicals and their combinations 73
Figure 16 Heatmap of chemical prevalence by product category for the DODSON chemicals 79
Figure 17 Heatmap showing prevalence by product category for the top-25 target chemicals 80
Figure 18 Number of products containing one or more DODSON chemicals 82
Figure 19 Tabulating per-product combinations of the target chemicals 83
Figure 20 Computing per-consumer EF and RF of the target chemicals and their combinations 87
Figure 21 Heatmap of prevalence by product category for the top-25 TOX21 chemicals 97
Figure 22 Heatmap of prevalence by product category for the top-25 HSDB chemicals 101
Figure 23 Heatmap of prevalence by product category for the top-25 CACOC chemicals 105
Figure 24 Heatmap of prevalence by product category for the top-25 EDCDB chemicals 109
Figure 25 Heatmap of prevalence by product category for the top-25 DODSON chemicals 113
Figure 26 Computational exposure science framework 134
Figure 27 Event-driven, asynchronous system to prioritize the target chemicals. 137

List of Tables

Table 1 Prevalence and synonymy of the DODSON chemicals in consumer products 15
Table 2 Preparing the target chemicals for matching to PubChem Compound..... 21
Table 3 Preprocessing chemical names 24
Table 4 UMLS vocabularies 25
Table 5 Extracting and cleaning data from the raw consumer product HTML files 37
Table 6 Assigning product categories 48
Table 7 Product sample size for the original CPDB (Gabb and Blake, 2016a) 50
Table 8 Breakdown of personal care products by category in the Kantar dataset 52
Table 9 Mapping the target chemicals to PubChem CIDs 55
Table 10 Final breakdown of the authoritative lists of target chemicals 57
Table 11 Tabulating target chemical proportions by product category..... 61
Table 12 Manual validation of ingredient string matching to PubChem 63
Table 13 Manual analysis of unmatched ingredient strings..... 64
Table 14 Homonymy among the DODSON chemicals (taken from Gabb and Blake, 2016a)..... 70
Table 15 Twenty-five most prevalent target chemicals in the complete product sample 77
Table 16 Tabulating per-product combinations of the target chemicals..... 84
Table 17 Twenty most common 2-way per-product chemical combinations..... 85
Table 18 Twenty most common 3-way per-product chemical combinations..... 86
Table 19 Tabulating per-consumer combinations of the target chemicals..... 88
Table 20 Top-20 2-way per-consumer chemical combinations ranked by EF..... 90
Table 21 Top-20 3-way per-consumer chemical combinations ranked by EF..... 91
Table 22 Top-20 2-way per-consumer chemical combinations ranked by RF..... 92
Table 23 Top-20 3-way per-consumer chemical combinations ranked by RF..... 93
Table 24 Comparison of the top-25 chemicals ranked using four approaches..... 95
Table 25 Top-25 2-way TOX21 chemical combinations ranked by RF..... 98
Table 26 Top-25 3-way TOX21 chemical combinations ranked by RF..... 99
Table 27 Top-25 2-way HSDB chemical combinations ranked by RF 102
Table 28 Top-25 3-way HSDB chemical combinations ranked by RF 103

Table 29 Top-25 2-way CACOC chemical combinations ranked by RF 106
Table 30 Top-25 3-way CACOC chemical combinations ranked by RF 107
Table 31 Top-25 2-way EDCDB chemical combinations ranked by RF 110
Table 32 Top-25 3-way EDCDB chemical combinations ranked by RF 111
Table 33 Top-25 2-way DODSON chemical combinations ranked by RF 114
Table 34 Top-25 3-way DODSON chemical combinations ranked by RF 115
Table 35 Comparison of the top-25 chemicals from each authoritative list ranked by RF 117

Chapter 1: Introduction and Motivation

The United States and Europe have opposite regulatory approaches to chemical usage (GAO, 2007). Under regulation EC 1907/2006 (Registering, Evaluation, Authorization and Restriction of Chemicals), European manufacturers must certify that the chemical ingredients in their products are safe. Under the U.S. Toxic Substances Control Act of 1976, the burden is on the EPA to demonstrate potential harm before imposing regulations on manufacturers. Manufacturers are under no legal obligation to perform toxicological analysis on chemicals that have already been approved for import and use. However, approval for import and use does not guarantee safety, especially for long-term exposure.

Roughly 7.9 million chemicals are currently available for purchase (Chuprina et al., 2010) and 80,000 chemicals are currently registered under the U.S. Toxic Substances Control Act of 1976 (TCSA, 1976). The potential risk to humans from exposure to these chemicals has been recognized for decades (Bracken and Weiss, 1977, p. 203):

> *"Initially, when considering the toxicity of an individual product, the concentration of a chemical may appear to be innocuous. However, when evaluated from a global viewpoint with knowledge of total exposure to the chemicals, there may be indications that the safe threshold level for that chemical has been surpassed. The consumer is exposed to cumulative toxic effects of chemicals or products and to unknown synergistic or antagonistic effects resulting from constant or frequent exposure to chemicals presently unidentified in consumer products."*

This has become an accepted—or perhaps ignored—tradeoff of life in modern society. However, Christopher Wild, a cancer epidemiologist, recognized that current disease trends cannot be explained by genetics alone. For example, autism, asthma, and leukemia are on the rise (Perrin et al., 2007; Hertz-Picciotto and Delwiche, 2009; Meeker, 2012), and biomonitoring studies reveal widespread exposure to environmental chemicals (Becker et al., 2007; CDC, 2011; Park et al., 2012). Wild suggested that the "exposome" (the combined exposures over one's lifetime) be considered alongside the genome (Wild, 2005, 2012). The rationale is well-summarized by Dennis et al. (2016, p. 1505):

> *"Exogenous chemicals can cause thousands of perturbations to our bodies. However, from a health standpoint, we are most concerned with those effects that*

> *are most likely to disrupt our health. It is rather amazing that faced with altered temperature, activity, energy uptake, and psychological challenges, we can maintain a rather consistent blood pressure, weight, and body temperature. These key functions operate under a series of cooperative homeostatic mechanisms that sense alterations and respond in a way to minimize the change in the system. However, the goal of these systems is not always to return the system to exactly where it was before the challenge. This process of dynamic homeostasis has been termed allostasis, with the concept of allostatic load representing the cost of the cumulative correcting process. By capturing the wear and tear process on our bodies, allostatic load may provide a clinically relevant means of measuring the biological response as it relates to the exposome."*

The "Swiss cheese model of adverse effects" (Boekelheide and Campion, 2010) illustrates allostatic load graphically, in which a series of latent chemical exposure effects combine to ultimately disrupt health (Figure 1). Developing fetuses and children are particularly vulnerable (Rice and Barone, 2000; Selevan et al., 2000) because exposure during development can cause lifelong, permanent effects (Palanza et al., 1999; Welshons et al., 2006). Epigenetic changes could even affect future generations (Perera and Herbstman, 2011; Ho et al., 2012).

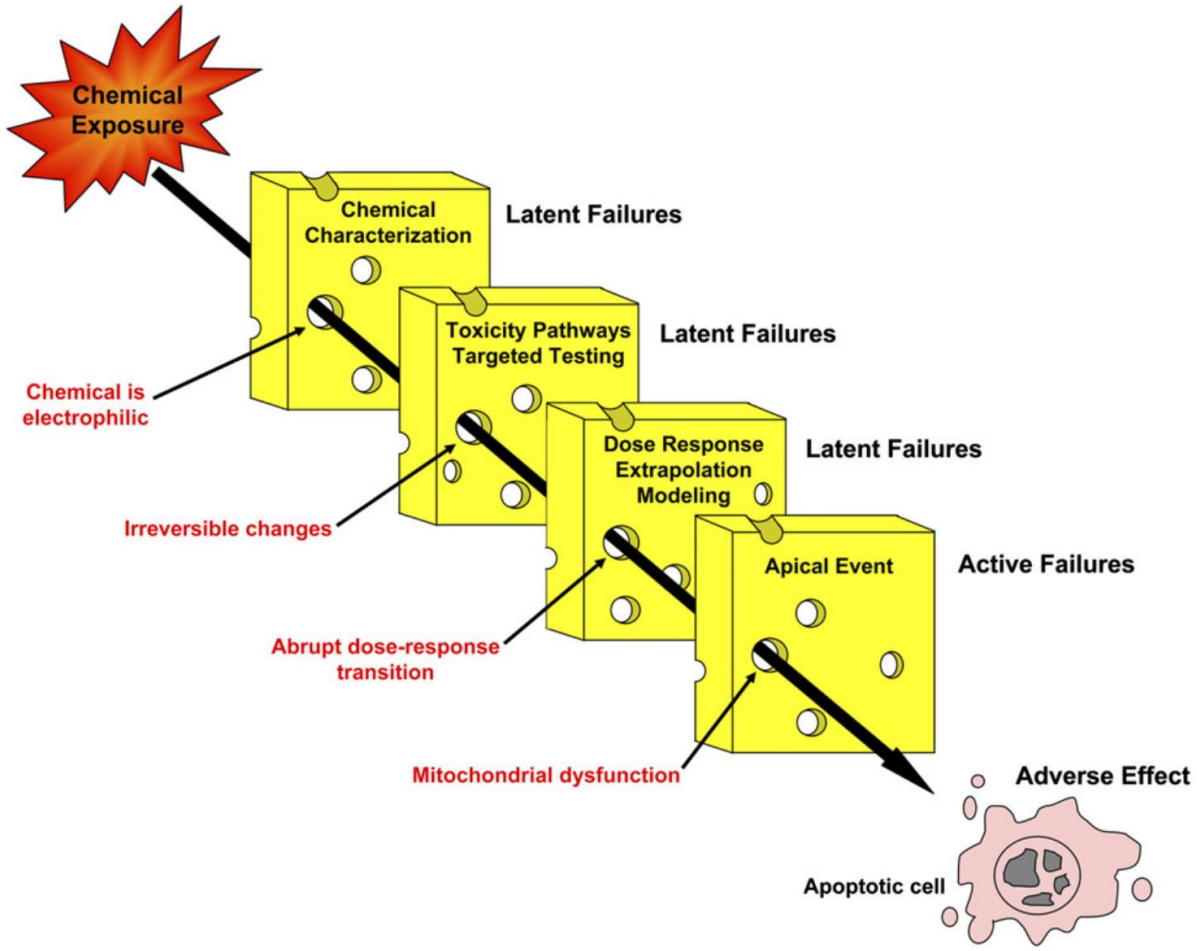


Figure 1 Swiss cheese model of adverse effects

Taken from Boekelheide and Campion (2010), this model postulates that the latent failures from chemical exposure(s) can eventually result in active failure. In terms of allostatic load, each latent failure is in a state of dynamic homeostasis, or allostasis. Disease results when the allostatic load becomes too great and the biochemical system can no longer achieve homeostasis.

Given the number of chemicals in widespread use, it is infeasible to perform toxicity screening and risk assessment on all of them. The Toxicology Testing in the 21st Century (TOX21) program – a collaboration among U.S. federal agencies including the Environmental Protection Agency (EPA), the National Institutes of Health (NIH), and the Food and Drug Administration (FDA) – compiled a list of 10,000+ chemicals (called the TOX21 10K Library, containing 9,011 unique chemicals) that have the "potential to disrupt biological pathways that may result in toxicity" (EPA, 2008; NRC, 2007; Tice et al., 2013). These chemicals are in widespread use, but most have not been subjected to thorough toxicological screening (Dix et al., 2007; Sheldon and Cohen Hubal, 2009). To date, the EPA Integrated Risk Information

System contains risk assessments for only 511 chemicals (IRIS, 2017). The EPA goal is to perform risk assessment on the entire TOX21 10K Library. Screening and risk assessment are expensive, so intelligent approaches to prioritization are needed to focus toxicology research on chemicals that have the greatest potential impact on consumer health (Boekelheide and Campion, 2010; Krewski et al., 2014).

Much of the work in assessing risks associated with chemical exposure focuses on individual chemicals. However, communities face exposure from a variety of sources, and body burden (or chemical load) is significantly higher than a century ago (Glegg and Richards, 2007; Greggs et al., 2013; Sanderson et al., 2013). More importantly, the dose-response for chemical mixtures may be independent (additive), synergistic, or antagonistic (Sexton and Hattis, 2007; Pollock et al., 2017) and health outcomes can be influenced by both chemical and non-chemical stressors. With respect to chemicals, far-field exposure (i.e., the aggregate environmental intake of a chemical) from persistent, high-production-volume (HPV) chemicals (defined by the EPA as those with an annual U.S. production or importation greater than one million pounds) has been well-explored (Muir and Howard, 2006), but near-field exposure from everyday consumer products such as shampoo, toothpaste, and makeup accounts for a significant portion of our overall chemical load (Dodson et al., 2012; Egeghy et al., 2011; Koniecki et al., 2011). Chemicals from consumer product usage have been detected in blood and urine (Wambaugh et al., 2013; Harley et al., 2016). Also, near-field chemical exposure from consumer products is generally larger than the doses resulting from far-field industrial exposure sources (Ott, 1990; Wallace, 1991).

Consumer products contain ingredients that can be beneficial or harmful depending on their concentration and co-exposure to other environmental chemicals. Thousands of different chemical ingredients are used in consumer products. Gabb and Blake (2016a) identified 7,486 distinct chemicals in a sample of 38,975 consumer products. Recognition of the potential risk prompted the Consumer Product Safety Act of 1972 (CPSA) and the creation of the Consumer Product Safety Commission (CPSC). However, its authority to regulate the chemical ingredients in consumer products is limited.

In response to this increased awareness, risk assessments that once focused on a single pesticide or chemical (e.g., benzene, dioxin, and PCBs) are moving toward a less isolated and better contextualized view of the multiple environmental agents to which humans are exposed (Jayjock et al., 2009). Cumulative risk assessments (CRA) consider multiple chemical and environmental stressors, though there is no single approach to measuring exposure (Choudhury et al., 2000; EPA, 1986). The most challenging type of chemical mixtures to assess are the so-called "coincidental mixtures" that "occur by happenstance at a time or place of interest" (Sexton and Hattis, 2007). It is not feasible to test every possible chemical mixture, so new methods are needed to prioritize based on the level of human exposure (Dix et al., 2007; Sheldon and Cohen Hubal, 2009), the nature of exposure, the severity of effects, and likelihood of interactions (Sexton and Hattis, 2007).

Endocrine disrupting compounds (EDCs), which are chemicals that mimic hormones and alter endocrine signaling, are of particular interest because of their subtle and potentially far-reaching health effects (Colborn et al., 1993; Crisp et al., 1998; WHO/UNEP, 2013), including effects on oncogenesis (Soto and Sonnenschein, 2010), metabolism (Elobeid and Allison, 2008; Grun and Blumberg, 2009; Heindel, 2003; Newbold, 2010; Newbold et al., 2008; Goodman et al., 2014), reproductive and nervous system development (Hengstler et al., 2011; Ejaredar et al., 2015), and reproductive health (Pollack et al., 2018). Epidemiological studies have reported associations between prenatal exposure to chemicals classified as EDCs and early cognitive development (Engel et al., 2010; Factor-Litvak et al., 2014). In addition to potential health effects that may be subtle and difficult to observe, EDCs have also been associated with conditions like asthma. For example, some fragrance compounds may act as direct irritants to exacerbate and perhaps even cause asthma and other respiratory disorders (Bridges, 2002; Kumar et al., 1995). In addition, there is evidence that some EDCs, including triclosan, glycol ethers, and phthalates can exacerbate asthma indirectly via immune sensitization (Anderson et al., 2013; Bornehag and Nanberg, 2010; Bornehag et al., 2004; Choi et al., 2010; North et al., 2014).

Informatics approaches can assist in prioritizing chemicals for CRA by integrating data from multiple sources (Jayjock et al., 2009; Sheldon and Cohen Hubal, 2009). For example, the

EPA's NexGen risk assessment framework explored a range of methods, including rapid screening to prioritize potentially harmful chemicals (Cohen Hubal et al., 2010; Collins et al., 2008; Cote et al., 2012; Dix et al., 2007; Egeghy et al., 2011; Krewski et al., 2014). The goal of the present study is to help prioritize individual and chemical combinations based on near-field exposure from everyday consumer products, which accounts for a significant portion of overall body burden (Dodson et al., 2012; Egeghy et al., 2011; Koniecki et al., 2011). The emphasis on such products is motivated in part by the frequency and type of exposure (consider products such as deodorant or toothpaste that are used every day and are applied directly to the skin or mucosa). In contrast to some environmental exposures where either community or regulatory pressure is needed to change exposure levels, individual consumers have more control over the products that they use, and hence their exposure levels.

This control is not absolute, however. Some consumer products (e.g., vinyl shower curtains and pillow protectors, plastic storage containers) do not typically provide an ingredient list but may contain potentially harmful plasticizers (Dodson et al., 2012). When an ingredient list is provided, fragrance and flavoring chemicals are sometimes listed as generic "fragrance" or "flavor." Fragrance and flavor mixtures can be designated trade secrets under the Fair Packaging and Labeling Act of 1967 (FPLA, § 1454.c.3.B) so their chemical composition need not be divulged. Also, plasticizers leached into a product from the container are not listed (Erythropel et al., 2014; Yang et al., 2011). Also, there may simply be a lack of safer alternative ingredients for consumers to choose. Finally, chemical synonymy, or different names referring to the same chemical, adds a layer of obfuscation that can hinder consumer identification of potentially harmful ingredients. The FPLA was a good step toward empowering consumers to make informed decisions about the products that they use. However, incomplete and confusing product labels undermine the informed consent that the FPLA attempts to provide. Based on the results in Gabb and Blake (2016a), a case can be made to amend the FPLA to standardize ingredient nomenclature, particularly with respect to harmful or potentially harmful ingredients like those identified by TOX21 or the California Department of Toxic Substances Control.

The present research develops and evaluates a data-driven approach to prioritizing the risk assessment of potentially harmful chemical ingredients (i.e., the aggregate exposure to

individual chemicals) and ingredient combinations (i.e., the cumulative, simultaneous exposure to multiple chemicals) based on their prevalence in everyday consumer products and the daily product usage patterns of typical consumers (Figure 2). There are thousands of chemicals in widespread use that have not been subjected to CRA. Government agencies responsible for consumer and environmental safety need intelligent, evidence-based prioritization of potentially harmful chemicals because CRA is time-consuming and expensive and their limited resources must be allocated efficiently. The present research prioritizes various authoritative lists of potentially harmful chemicals based on near-field exposure from consumer products. Such prioritization will help responsible agencies efficiently allocate resources for screening chemicals to which consumers are routinely exposed. Subsidiary benefits of the proposed research include assessments of how chemical synonymy undermines the informed consent that the FPLA ostensibly provides and analysis of the authoritative lists of potentially harmful chemicals. It may be the case that some authoritative lists are more appropriate than others in the context of consumer product usage.

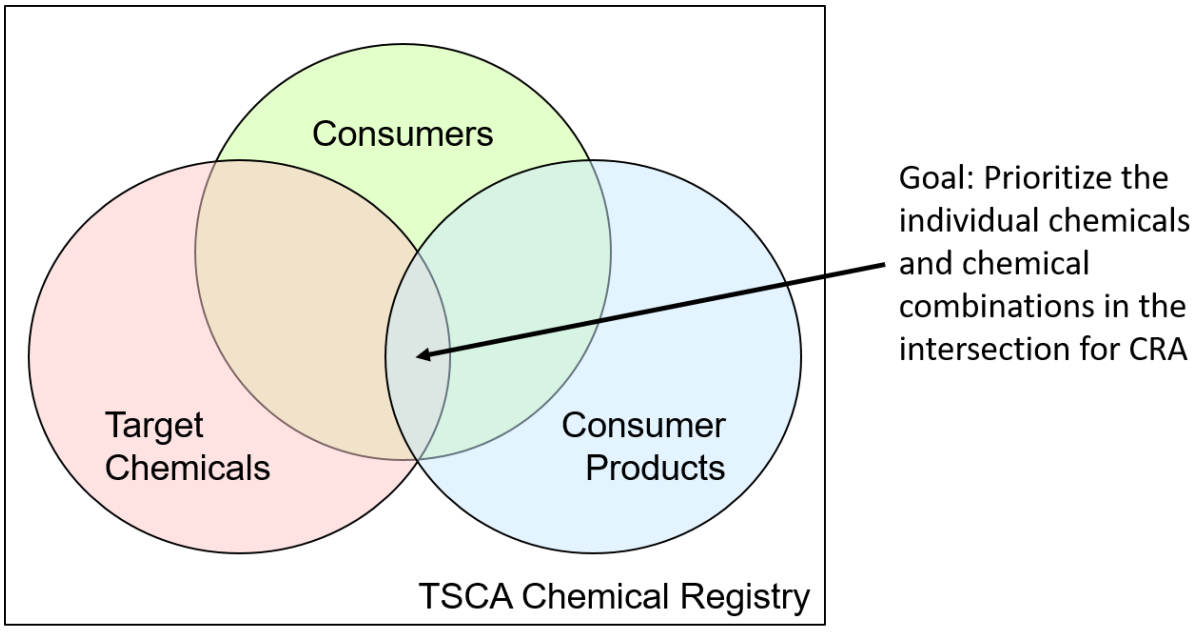


Figure 2 Research goal

Within the overall set of chemicals in the TSCA Chemical Registry, there exists of subset of target chemicals that occur in products used by consumers. The primary goal of the present research is to prioritize these chemicals and their combinations for CRA based on near-field exposure among consumers.

This dissertation is organized as follows. Chapter 2 reviews prior research that is related to the present study; namely, the creation of consumer product databases and alternative approaches to prioritizing chemicals for risk assessment. Chapters 3 and 4 describe the steps to gather and process the data necessary to conduct the research described herein, and much of the rationale behind these processes. Chapter 3 deals with the data collection: selecting the target chemicals to prioritize, choosing the chemical dictionary with which to unambiguously match target chemicals to consumer product ingredients, scraping consumer product information from an online retailer, and parsing this data into structured formats. Chapter 4 describes the data cleaning processes and various exploratory analyses and evaluations of the final datasets. Source code described in these chapters is released in the spirit of Barnes (2010): "That the code is a little raw is one of the main reasons scientists give for not sharing it with others. Yet, software in all trades is written to be good enough for the job intended." The software for this research is no different. It is extensively commented but users are expected to be proficient in Python and regular expressions, and to a lesser extent in Java, awk, sed, and various Linux command-line utilities. Data are made available in the spirit of reproducibility outlined by Stodden et al. (2010, 2013) except where release would violate copyright (i.e., raw HTML scraped from online retailers) or terms of use (i.e., data purchased from Kantar Worldpanel). Chapter 5 describes the computation of the chemical exposure and retention factors that are used to prioritize the target chemicals. Chapters 6 and 7 present the analyses and discuss the results. Conclusions and ideas for future research are provided in the final chapter.

Chapter 2: Related Work

2.1 Generating Consumer Product Databases

Determining chemical prevalence in consumer products requires a comprehensive database of products and their ingredients. The idea of building databases of consumer products and their ingredients is not new, nor is the use of information systems to resolve synonymy among chemical ingredients. The CPSC, the government agency "responsible for programs that reduce the hazard of human injury from chemical consumer products," published regulatory guidelines over 40 years ago "for obtaining chemical formulation information for specified consumer products" (Bracken and Weiss, 1977, p. 202). The CPSC attempted to compile a database of consumer products and their ingredients (Byer et al., 1976) but this effort initially stalled due to objections from manufacturers and trade associations that the Commission was creating undue burden and overstepping its legal authority under the CPSA. These legal objections were eventually resolved through further legislation, but this early CPSC effort represents the first attempt at a data-driven approach to the chemical safety of consumer products. The importance of accounting for synonymy among ingredient names was also recognized in this early work (Landau and Byer, 1976; Tate, 1967).

Product labeling requirements have improved since the CPSC attempted to compile its database (Byer et al., 1976). For example, products regulated by the FDA (mainly cosmetics and personal care products) have several requirements (FDA, 2017). First, all ingredients that are *intentionally* added to the product must be disclosed. Chemicals leached from product packaging, chemical degradation byproducts, etc. need not be disclosed. Second, ingredients must be listed in descending order of prevalence; more specifically, in descending order of weight fraction. In the case of medications, active ingredients (i.e., those included for a specific biological activity) and their exact weight fractions must be listed first. If weight fraction is below 0.01, ingredients can be listed in any order. Colorants can be listed in any order at the end of the ingredient list. Finally, the chemicals in fragrance and flavor mixtures can be listed explicitly or designated simply as generic "fragrance" or "flavor" in the ingredient list. As mentioned previously, such mixtures can be designated trade secrets under the Fair Packaging and Labeling Act of 1967 (FPLA, § 1454.c.3.B) so their chemical composition need not be

divulged. Though current labeling regulations do not require complete disclosure of the product formulation, and may differ for classes of consumer products (i.e., household and automotive) that are regulated by other Government agencies, the availability of product formulation data is generally better than it once was.

More modern databases have been compiled and made available to varying degrees, e.g.: Skin Deep (http://www.ewg.org/skindeep/) created by the Environmental Working Group and the Consumer Product Information Database (http://whatsinproducts.com/pages/index/1) created by DeLima Associates. While these databases are comprehensive, neither is freely downloadable or otherwise amenable to bulk querying or integration with other data sources. The database at the center of the present research, the Consumer Products Database (CPDB), is compiled by scraping publicly available data from online consumer product retailers (Gabb and Blake, 2016a). It was created out of necessity because an off-the-shelf database that could be installed locally was not available. Goldsmith et al. (2014) used a similar approach to compile a consumer product database, but they scraped Material Safety Data Sheets (MSDS) rather than the product pages of online retailers. MSDS are a good but incomplete source of product data. Unlike a product label, an MSDS is only required to list ingredients that are known to be hazardous. Therefore, many potentially harmful chemicals do not appear in MSDS, as noted in Gabb and Blake (2016a). The present research focuses on chemical ingredients that are suspected of being hazardous but have not yet been thoroughly screened for toxicity. Consequently, the CPDB is much larger and more comprehensive than the database described in Goldsmith et al. (2014). The CPDB contains 38,975 products with 32,231 distinct ingredient names, while Goldsmith et al.'s (2014) database only contains 8,921 products with 1,797 unique ingredients. As there are 9,011 chemicals in the TOX21 library alone, the MSDS approach is unlikely to be useful for prioritizing potentially hazardous chemicals for risk assessment.

2.2 Prioritizing Chemicals for Risk Assessment

There are many approaches to prioritization, each with advantages and disadvantages. For example, production volume is an objective proxy for potential far-field chemical exposure (i.e., the aggregate environmental intake of a chemical) (Muir and Howard, 2006). The EPA

provides several computational toxicology resources to assist with risk-based chemical prioritization: the Toxicity Reference Database (ToxRefDB; Knudsen et al., 2009; Martin et al., 2009a; Martin et al., 2009b), the Distributed Structure-Searchable Toxicity Database (DSSTox; Richard and Williams, 2002; Richard et al., 2006), the Toxicity Forecaster (ToxCast; Dix et al., 2007; Richard et al., 2016), and the Exposure Forecaster (ExpoCast; Cohen Hubal et al., 2010; Judson et al., 2012; Wambaugh et al., 2013). The Aggregated Computational Toxicology Resource (ACToR) aggregates these databases under one interface (Judson et al., 2012). ToxRefDB compiles the results from published animal pesticide assays. The TOX21 program is moving away from *in vivo* testing in favor of *in vitro* high-throughput toxicity screening (HTS), so the ToxRefDB is mostly a legacy project even though it continues to provide useful data to bench toxicologists. DSSTox is a database of chemical structures and associated toxicity annotations that aims to prioritize the screening of potentially harmful chemicals based on structure-activity relationship (SAR) modeling. ToxCast prioritizes chemicals for risk assessment based on HTS, in which isolated cells and proteins are exposed to chemicals. DSSTox and ToxCast provide useful and objective information about potential toxicity but neither takes likely exposure or retention into account, nor do they consider combined exposure. ExpoCast is an exposure-based prioritization framework. Its current models are based on multiple empirical analyses of indoor and outdoor air, drinking water, soil, urine, dust, and indoor surface residues for selected chemicals, mainly pesticides and their metabolites. Consumer products are not considered, so ExpoCast is not directly useful for the present research, but the results of this research could supplement the ExpoCast exposure forecasting models.

There are many approaches to estimating chemical exposure in the presence or absence of biomonitoring data (e.g., blood tests and urinalyses). The National Health and Nutrition Examination Survey (NHANES) from the Centers for Disease Control and Prevention (CDC) is the primary source of biomonitoring data (CDC, 2011). Wambaugh et al. (2014) inferred 106 parent chemicals from NHANES urine metabolite data and derived five predictive characteristics from these chemicals: high production volume, active or inert pesticide ingredient, industrial use, and use in consumer products. Their heuristic model was used to prioritize the TOX21 chemicals, but far-field characteristics dominate the model. Sanderson et al. (2006, 2013)

further provide a framework to estimate human exposure and risk that includes exposure from consumer products. They selected 291 HPV chemicals and divided them into 10 broad classes (e.g., amine oxides, aliphatic alcohols) that were cross-indexed with various consumer product categories to provide concentration estimates to prioritize risk assessment. Their framework takes consumer products into account when estimating exposure, but does not look at specific chemicals or chemical combinations or take actual consumer usage into account. It simply uses aggregated data from manufacturer surveys.

"Exposure models can be used to estimate exposures to chemicals in the absence of biomonitoring data and as tools in chemical risk prioritization and screening" (Csiszar et al., 2017, p. 152). The latter study used a stochastic approach to model population variability and product usage mode (e.g., left on or rinsed off after application) and location (e.g., face, mouth, hair) to estimate paraben exposure from consumer products. It estimated product intake fractions from paraben-containing consumer product categories and converted them to urine levels for comparison with NHANES urinalysis data. However, the study had two flawed assumptions. First, it used average concentrations of methyl, ethyl, propyl, and butyl paraben by product category. The CPDB shows that products vary in their use of parabens, and the presence of one paraben compound does not necessarily mean that other parabens are also present (Gabb and Blake, 2016a). Second, toothpaste and mouthwash were excluded because "these products are not reported to contain parabens." Previous informatics and analytical studies did not find parabens in these products classes (Dodson et al., 2012; Goldsmith et al., 2014; Guo and Kannan, 2013). However, this is a false assumption resulting from analysis of consumer product samples that are too small. The larger CPDB sample includes toothpastes and mouthwashes that contain parabens.

Two previous studies have computed aggregate exposure (Cowan-Ellsberry and Robison, 2009; Comiskey et al., 2015). Comiskey et al. (2015) estimated aggregate exposure to generic fragrance (rather than specific fragrance chemicals) based on the Kantar Worldpanel consumer profiles (https://www.kantarworldpanel.com); the average fragrance content for a given product category; and a complex exposure model that accounts for typical usage amount, usage mode, usage location, and generalized dermal/oral absorption kinetics (Hall et al., 2007,

2011; Loretz et al., 2005, 2006, 2008; McNamara et al., 2007). Cowan-Ellsberry and Robison (2009) estimated the aggregate exposure to parabens (methyl, ethyl, propyl, and butyl paraben) from consumer products (mainly cosmetics) based on the average fractional concentration of each paraben for the product category, the average amount of product used, the average daily usage frequency, the estimated fractional retention of the product category (e.g., rinse-off factors for shampoos vs. toothpastes), and the body weight of the consumer. Their consumer usage data came from an internal Procter & Gamble survey of 3,297 American women. Cowan-Ellsberry and Robison (2009) improved over previous aggregate exposure estimations by considering specific chemicals as well as the possibility that a given product does not contain the chemical of interest.

Chapter 3: Data Collection

Achieving the goal of prioritizing the chemical ingredients in consumer product for CRA requires the integration of several datasets (Figure 2). First, a list of potentially harmful chemicals must be compiled. As noted previously, the number of potential targets is vast: roughly 7.9 million chemicals available for purchase (Chuprina et al., 2010) and 80,000 chemicals registered under the U.S. Toxic Substances Control Act of 1976 (TCSA, 1976). However, not all of them are toxicologically interesting so the list of targets is narrowed using the authoritative lists described in Chapter 3.1. Second, a chemical dictionary is needed to uniquely identify each target chemical and chemical ingredient. As noted previously, chemical nomenclature is imprecise. Correct identification of chemicals is absolutely required in order to determine prevalence in consumer products. The chemical dictionary used in this study and the reasons for its selection are described in Chapter 3.2. Third, in order to measure the prevalence of the target chemicals in consumer products, a database of consumer products and their formulations is obviously required. Chapter 3.3 describes the process to create such a database. Finally, incorporating consumer behavior into the prioritization scheme is a key goal of this analysis. The dataset of consumer product usage patterns is described in Chapter 3.4.

3.1 Selecting and Preprocessing the Target Chemicals

Two sets of target chemicals are used in the present analyses. The first set was selected from a prior gas chromatography-mass spectrometry (GCMS) analysis of 213 consumer products to measure the levels of 55 potential EDC and asthma-associated chemicals (Dodson et al., 2012). They are listed in Table 1. The biological effects of exposure to these chemicals are well-studied, making them a reasonable yet still manageable set for GCMS analysis and a good proof-of-concept for the informatics approach; namely, that the informatics approach can detect specific chemicals in a sample of consumer products scraped from online retailers (Gabb and Blake, 2016a). It is not an exhaustive set of potentially harmful chemicals, but it does provide a basis of comparison between the informatics approach and the prior GCMS analysis.

Table 1 Prevalence and synonymy of the DODSON chemicals in consumer products

Ingredient Class	Chemical Name	# Products Containing this Chemical	# Synonyms Appearing in Product Ingredient Lists	Synonyms (Number of Products)
UV filter	octinoxate	1287	4	octinoxate (556), octylmethoxycinnamate (30), octyl methoxycinnamate (46), ethylhexyl methoxycinnamate (655)
UV filter	benzophenone-3	450	2	oxybenzone (416), benzophenone-3 (34)
UV filter	benzophenone-1	0		
UV filter	benzophenone	5	1	benzophenone (5)
Cyclosiloxane	dodecamethylcyclohexasiloxane	0		
Cyclosiloxane	decamethylcyclopentasiloxane	625	2	decamethylcyclopentasiloxane (10), cyclomethicone (615)
Cyclosiloxane	octamethylcyclotetrasiloxane	7	1	octamethylcyclotetrasiloxane (7)
Glycol ether	2,2-butoxyethoxyethanol	3	1	butoxydiglycol (3)
Glycol ether	2,2-methoxyethoxyethanol	0		
Glycol ether	2-phenoxyethanol	5638	3	phenoxyethanol (5632), polyoxyethylene phenyl ether (1), 2 phenoxyethanol (5)
Glycol ether	2-butoxyethanol	5	2	butyl glycol (2), butoxyethanol (3)
Synthetic fragrance	phenethyl alcohol	193	4	phenethyl alcohol (180), phenylethyl alcohol (2), phenylethanol (6), phenyl ethyl alcohol (5)
Synthetic fragrance	musk xylene	0		
Synthetic fragrance	musk ketone	0		
Synthetic fragrance	methyl ionone	197	4	methyl ionone (6), alpha-isomethyl ionone (183), alpha-isomethylionone (5), methyl ionone gamma (3)
Synthetic fragrance	isobornyl acetate	1	1	bornyl acetate (1)
Synthetic fragrance	hhcb	0		
Synthetic fragrance	dpmi	0		
Synthetic fragrance	diphenyl ether	1	1	phenyl ether (1)
Synthetic fragrance	bucinal	539	2	lilial (71), butylphenyl methylpropional (468)

Table 1 (cont.)

Synthetic fragrance	ahtn	1	1	acetyl hexamethyl tetralin (1)
Natural fragrance	terpineol	4	2	terpineol (3), terpineol alpha (1)
Natural fragrance	pinene	0		
Natural fragrance	methyl salicylate	105	3	methyl salicylate (83), wintergreen oil (21), sweet birch oil (1)
Natural fragrance	methyl eugenol	0		
Natural fragrance	linalool	2517	2	linalool (2516), linalol (1)
Natural fragrance	limonene	2623	13	limonene (2334), d-limonene (17), limonen (1), orange flavor (44), lemon oil (83), lemon extract (15), sweet orange oil (4), orange oil (55), citrus limon oil (2), oil of lemon (2), orange flower oil (1), citrus sinensis oil (61), citrus sinensis peel oil (4)
Natural fragrance	hexyl cinnemal	56	4	hexyl cinnamic aldehyde (45), hexyl cinnamaldehyde (7), hexylcinnamaldehyde (3), alpha-hexylcinnamaldehyde (1)
Natural fragrance	eugenol	429	1	eugenol (429)
Natural fragrance	benzylacetate	0		
Alkylphenol	nonylphenol diethoxylate	0		
Alkylphenol	nonylphenol monoethoxylate	0		
Alkylphenol	4-t-nonylphenol	0		
Alkylphenol	octylphenol diethoxylate	0		
Alkylphenol	octylphenol monoethoxylate	29	4	octoxynol 9 (21), octoxynol-9 (3), octoxynol (1), octylphenoxypolyethoxyethanol (4)
Alkylphenol	4-t-octylphenol	0		
Ethanolamine	diethanolamine	16	1	diethanolamine (16)
Ethanolamine	monoethanolamine	97	2	ethanolamine (90), monoethanolamine (7)
Antimicrobial	triclosan	104	1	triclosan (104)
Antimicrobial	triclocarban	12	1	triclocarban (12)
Bisphenol A	bisphenol a	0		
Phthalate	diethyl phthalate	5	1	diethyl phthalate (5)
Phthalate	di-n-propyl phthalate	0		

Table 1 (cont.)

Phthalate	di-n-octyl phthalate	0		
Phthalate	di-n-hexyl phthalate	0		
Phthalate	di-n-butyl phthalate	26	1	dibutyl phthalate (26)
Phthalate	di-isononyl phthalate	0		
Phthalate	di-isobutyl phthalate	0		
Phthalate	di-cyclohexyl phthalate	0		
Phthalate	benzylbutyl phthalate	0		
Phthalate	bis(2-ethylhexyl) phthalate	0		
Phthalate	bis(2-ethylhexyl) adipate	29	2	diethylhexyl adipate (25), dioctyl adipate (4)
Paraben	butyl paraben	1015	2	butylparaben (1008), butyl paraben (7)
Paraben	ethyl paraben	1364	3	ethylparaben (1356), ethyl paraben (6), catalase (2)
Paraben	methyl paraben	4510	3	methylparaben (4435), methyl paraben (74), methyl 4-hydroxybenzoate (1)

The second set of target chemicals is used to demonstrate the scalability of the informatics approach to a much larger, but still toxicologically important, set of chemicals. Five authoritative lists of suspected harmful chemicals were selected: TOX21 (EPA, 2008; NRC, 2007; Tice et al., 2013), the Hazardous Substances Data Bank (HSDB; Fonger et al., 2000, 2014), the California Chemicals of Concern (CACOC; DTSC, 2016), the EDCs Data Bank (EDCDB; Montes-Grajales and Olivero-Verbel, 2015), and the original 55 chemicals (DODSON) from Dodson et al. (2012) that were used to demonstrate proof-of-concept for the informatics approach (Gabb and Blake, 2016a). These lists are manually curated by expert toxicologists and/or formal scientific review panels. The second, larger set of target chemicals consists of the union of TOX21, HSDB, CACOC, EDCDB, and DODSON.

TOX21 is suitable for this research for several reasons. First, it contains 9,011 unique chemicals so it is large enough to truly test the scalability of the informatics approach. Second, it was compiled and vetted by the TOX21 consortium, so it is scientifically reasonable. Third, the chemicals in this list have the "potential to disrupt biological pathways that may result in toxicity" (EPA, 2008; NRC, 2007; Tice et al., 2013) but have not been subjected to CRA so they are toxicologically interesting and there is a definite need for intelligent prioritization based on

exposure (Boekelheide and Campion, 2010; Dix et al., 2007; Krewski et al., 2014; Sexton and Hattis, 2007; Sheldon and Cohen Hubal, 2009). Fourth, the Kantar Worldpanel dataset described in Chapter 3.4 contains product usage patterns for American consumers and the CPDB contains products scraped from a U.S. retailer. It makes sense to use a list of chemicals that was compiled by U.S. agencies concerned with chemicals that are in widespread domestic use.

The Toxicology Data Network (Fonger et al., 2000, 2014), part of the National Library of Medicine (NLM), produces the HSDB (http://toxnet.nlm.nih.gov/newtoxnet/hsdb.htm), a list of 5,731 potentially hazardous chemicals. It is suitable for this research for the same reasons as TOX21: it is large, it is curated by a scientific review panel that meets several times yearly to update the database, and it focuses on chemicals that are potentially harmful.

The Safer Consumer Products program of the California Department of Toxic Substances Control publishes the CACOC, a list of 2,444 Candidate Chemicals (https://www.dtsc.ca.gov/SCP/CandidateChemicalsList.cfm) "that exhibit a hazard trait and/or an environmental or toxicological endpoint" (DTSC, 2016). Consumer products that contain one or more of these Candidate Chemicals are considered "Priority Products" that may be subject to Safer Consumer Product regulations. CACOC is suitable for this research for the same reasons as TOX21: it is large, it has been authoritatively vetted and is scientifically reasonable, and it targets potentially harmful chemicals with an eye toward American consumer products.

The EDCDB (http://edcs.unicartagena.edu.co/) contains 615 chemicals with potential endocrine disrupting effects (Montes-Grajales and Olivero-Verbel, 2015). This database incorporates the European Union and Endocrine Disruption Exchange lists of potential endocrine disruptors (EU, 2017; TEDX, 2017) (http://eng.mst.dk/chemicals/chemicals-in-products/endocrine-disruptors/the-eu-list-of-potential-endocrine-disruptors/ and https://endocrinedisruption.org/interactive-tools/tedx-list-of-potential-endocrine-disruptors/search-the-tedx-list). Though smaller than TOX21, HSDB, and CACOC, EDCDB is still suitable for the proposed research because it has been vetted and previous studies (Dodson et al., 2012; Gabb and Blake, 2016a) have confirmed the presence of EDCs from various chemical classes in consumer products.

Some of these authoritative lists are provided in relatively clean and ready-to-use format. Others require extensive preprocessing before they can be integrated into the informatics workflow. Preprocessing involves extracting the required, sometimes unstructured, data from these sources and putting them into a structured format. Figure 3 shows the various steps to process and combine the authoritative lists into a final set of target chemicals. The components of this workflow are described in Table 2. Preparing the list of target chemicals is largely a manual editing process but much of this editing has been automated using the Linux command-line utilities `sed`, `awk`, `grep`, `tr`, and `cat`. The sequence of editing commands to process each raw list of chemicals is placed in Linux shell scripts with extensive comments and instructions. These scripts are sensitive to document formats, so they should be used with caution if new lists are downloaded from the primary sources.

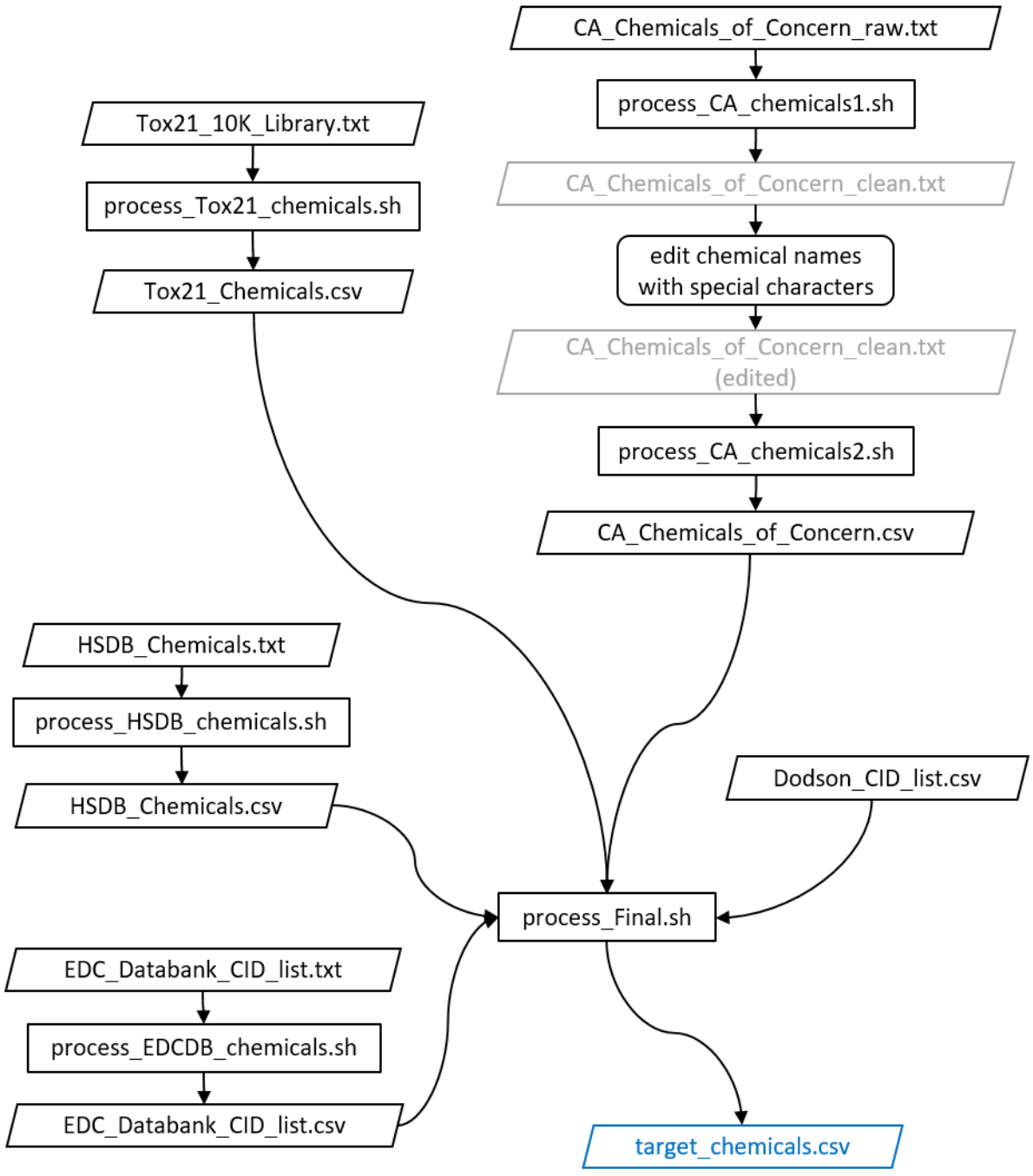


Figure 3 Preprocessing the authoritative lists of target chemicals

Rectangles indicate computational processes, rounded rectangles indicate manual processes, parallelograms indicate files, and arrows indicate data flow and dependencies. Blue parallelograms indicate data used in subsequent stages. Gray parallelograms indicate intermediate or validation data that are not used in subsequent stages.

Table 2 Preparing the target chemicals for matching to PubChem Compound

Computational Processes	
File Name	**Description**
process_Tox21_chemicals.sh, process_HSDB_chemicals.sh, process_EDCDB_chemicals.sh	These shell scripts process their respective authoritative lists into a common structured format.

`Command: process_Tox21_chemicals.sh`
`process_HSDB_chemicals.sh`
`process_EDCDB_chemicals.sh`	

process_CA_chemicals1.sh, process_CA_chemicals2.sh	These shell scripts process the unstructured CA Chemicals of Concern list into a common structure. Manual editing to resolve chemical names that contain special characters (mostly Greek letters) is required between the first and second scripts.
`Command: process_CA_chemicals1.sh` `process_CA_chemicals2.sh`	
process_Final.sh	This shell script performs additional processing to remove extraneous text from the chemical names (mostly trailing parenthetical information) and concatenates the preprocessed authoritative lists into the final list of target chemicals.
`Command: process_Final.sh`	
Data	
File Name	**Description**
Tox21_10K_Library.txt	This tab-delimited file from https://ntp.niehs.nih.gov/results/tox21/index.html contains a chemical name, chemical code (e.g., CAS-RN), and description in each record. The description indicates whether the chemical is a single compound, a mixture of stereoisomers, or a mixture/formulation. The name and code are used to positively identify the chemical. The description is not used in this analysis.
CA_Chemicals_of_Concern_raw.txt	This file contains the raw, unstructured list from https://calsafer.dtsc.ca.gov/chemical/search.aspx. Though unstructured, it contains markup tags that are used by `process_CA_chemicals1.sh` to extract the name and CAS-RN of each chemical in the list.

Table 2 (cont.)

CA_Chemicals_of_Concern_clean.txt	Each record in this pipe-delimited file contains the name, CAS-RN, and various other toxicological information for the CA Chemicals of Concern. Only the chemical name and CAS-RN are used in this analysis.
HSDB_Chemicals.txt	Each record in this tab-delimited file from https://www.nlm.nih.gov/databases/download/hsdb.html contains a chemical name, a CAS-RN, and the date when the chemical was added to the list. The latter is not used in this analysis.
EDC_Databank_CID_list.txt	This file is simply a list of PubChem Compound Identifiers (CID) downloaded from http://edcs.unicartagena.edu.co.
Tox21_Chemicals.csv, HSDB_Chemicals.csv, CA_Chemicals_of_Concern.csv, EDC_Databank_CID_list.csv, Dodson_CID_list.csv	These tab-delimited files contain the preprocessed data from each authoritative list of potentially harmful chemicals. Each record consists of four fields: an arbitrary serial number, a list identifier (TOX21, HSDB, CACOC, EDCDB, or DODSON), a CID (if provided), and a chemical name or CAS-RN (if provided). Not every field is populated at this stage.
target_chemicals.csv	This tab-delimited file contains the final, preprocessed list of all target chemicals from the authoritative lists. Each record consists of four fields: an arbitrary serial number, a list identifier, a CID, and a chemical name or CAS-RN. Not every field is populated at this stage.

Computational processes and files to prepare the authoritative lists of target chemicals for matching against PubChem Compound. Backslashes indicate command-line continuation.

3.2 Selecting the Chemical Dictionary

PubChem was used to assign unique identifiers to chemicals and to unify synonymous chemical names. PubChem was launched in 2004 as a repository of information about the biological activity of small molecules. It is hosted by the National Center for Biotechnology Information (NCBI). “The primary aim of PubChem is to provide a public on-line resource of comprehensive information on the biological activities of small molecules accessible to molecular biologists as well as computation and medicinal chemists” (Bolton et al., 2008). It consists of three distinct, community-supported databases: PubChem Substance, PubChem Compound, and PubChem BioAssay that are interlinked through substance, compound, and

assay identifiers. Users contribute and validate data but the actual PubChem database processing is highly automated and there is little manual curation or central control of input by the NCBI (Bolton et al., 2008).

The PubChem Compound (Kim et al., 2016) database is more appropriate for the purpose of matching product ingredient names to chemical identifiers because its chemical synonym list is large and it generally maps chemicals to Chemical Abstracts Service Registry Numbers (CAS-RN) and IUPAC International Chemical Identifiers (InChI). It also maps chemicals to Medical Subject Headings (MeSH) to facilitate integration with PubMed and the Unified Medical Language System (UMLS). The lists of synonyms for each CID were downloaded from PubChem in August 2016: ftp://ftp.ncbi.nlm.nih.gov/pubchem/compound/extras/cid-synonym-filtered.gz and ftp://ftp.ncbi.nlm.nih.gov/pubchem/compound/extras/cid-mesh. This file contained approximately 39 million CIDs and 150 million synonyms.

Some preprocessing is required to optimize name matching (Table 3). The transformations used here are similar to those applied to other chemical dictionaries and chemistry text processing applications (Hettne et al., 2009; McCray et al., 2001; Rogers and Aronson, 2008; Schwartz and Hearst, 2003). First, each synonym is converted to lowercase. Second, the long and abbreviated forms of a synonym [e.g., “acetyl hexamethyl tetralin (ahtn)”] are separated. The trailing, parenthetical text becomes a new synonym. Third, syntactic inversion is performed on synonyms that contain a comma followed by a space. For example, acetyl hexamethyl tetralin has a synonym “ethanone,␣1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-” (the ␣ symbol denotes a space) that is inverted to yield an additional synonym “1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-ethanone.” Fourth, some PubChem synonyms contain trailing, square-bracketed descriptors that are not part of the chemical name. For example, “mepacrine [inn:ban]” indicates that mepacrine is an International Nonproprietary Name and a British Approved Name. Adding another synonym without the bracketed text could open more matching possibilities. Fifth, most punctuation in systematic names serves a purpose but dashes and whitespace can be ignored during name matching. For example, “methyl␣paraben” and “methylparaben” are the same chemical, as are

"2-phenoxyethanol" and "2␣phenoxyethanol," so ignoring dashes and spaces would allow matches in cases where a particular permutation is missing from PubChem. Finally, synonyms shorter than three characters are discarded. Any duplicate synonyms resulting from preprocessing are also discarded.

Table 3 Preprocessing chemical names

Original PubChem Synonym	Dictionary Entries after Preprocessing	Preprocessing Applied to Synonym
mepacrine [inn:ban]	mepacrine [inn:ban]	None
	mepacrine	Trailing descriptor removed
acetyl hexamethyl tetralin (ahtn)	acetyl hexamethyl tetralin (ahtn)	None
	acetyl hexamethyl tetralin	Long form, abbreviation removed
	ahtn	Abbreviation added to dictionary
	acetylhexamethyltetralin	Spaces removed
ethanone, 1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-	ethanone, 1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-	None
	1-(5,6,7,8-tetrahydro-3,5,5,6,8,8-hexamethyl-2-naphthalenyl)-ethanone	Syntactic inversion
	ethanone,1(5,6,7,8tetrahydro3,5,5,6,8,8hexamethyl2naphthalenyl)	Spaces and dashes removed
	1(5,6,7,8tetrahydro3,5,5,6,8,8hexamethyl2naphthalenyl)ethanone	Syntactic inversion, spaces and dashes removed
methyl paraben	methyl paraben	None
	methylparaben	Space removed
2-phenoxyethanol	2-phenoxyethanol	None
	2phenoxyethanol	Dash removed

The UMLS was used in the early stages of this project to supplement PubChem. The UMLS project began in 1986 at the National Library of Medicine and the first version was released in 1989 (Humphreys and Lindberg, 1993; Humphreys et al., 1998). The UMLS is comprised of three components: the SPECIALIST lexicon, a semantic network, and a metathesaurus that aligns the content of 170 different independently maintained controlled vocabularies covering many aspects of biomedicine (e.g., diseases, drug and chemicals, surgical procedures, literature indexing, medical billing). A controlled vocabulary is a curated list of terms that represent the important concepts of a particular field. The terms in these

vocabularies are mapped to Concept Unique Identifiers (CUI). The UMLS was downloaded from http://www.nlm.nih.gov/research/umls in December 2014. Fifteen vocabularies were included and the number of terms in each vocabulary gives its relative contribution to the UMLS installation (Table 4). The strings associated with each concept undergo preprocessing similar to that of Hettne et al. (2010) to obtain a list of terms that can be matched to product ingredient names.

Table 4 UMLS vocabularies

Vocabulary	# Terms	Official Name
AOD	20,685	Alcohol and Other Drug Thesaurus
CHV	146,324	Consumer Health Vocabulary
DXP	10,113	DXplain (an expert diagnosis program)
MSH	815,608	Medical Subject Headings
MTH	171,407	UMLS Metathesaurus
MTHFDA	86,069	Metathesaurus FDA National Drug Code Directory
MTHSPL	113,248	Metathesaurus FDA Structured Product Labels
NCBI	1,265,703	National Center for Biotechnology Information Taxonomy
NCI	255,108	National Cancer Institute Thesaurus
RXNORM	628,521	RxNorm Vocabulary
SNM	44,274	Systemized Nomenclature of Medicine
SNMI	164,179	Systemized Nomenclature of Human and Veterinary Medicine
SNOMEDCT_US	1,225,189	Systemized Nomenclature of Medicine – Clinical Terms (U.S. Edition)
SNOMEDCT_VET	89,572	Veterinary Extension to SNOMED-CT
SRC	1,018	Metathesaurus Source Terminology Names

These vocabularies were used in Gabb and Blake (2016a). A vocabulary is a curated list of terms that represent the important concepts of a particular field. The number of terms in each vocabulary gives its relative contribution to the UMLS installation.

Synonyms must resolve to the same identifier if they are to be useful. In the UMLS, this identifier is the CUI. For example, searching the UMLS for octinoxate, octylmethoxycinnamate, octyl methoxycinnamate, or ethylhexyl methoxycinnamate returns the same CUI (C0046100). Searching the UMLS for C0046100 will return octinoxate and all of its synonyms. PubChem

performs the same function but refers to its unique identifiers as CIDs. Octinoxate, octylmethoxycinnamate, octyl methoxycinnamate, and ethylhexyl methoxycinnamate all have the same CID (5355130). Searching PubChem for 5355130 will return octinoxate and all of its synonyms.

The UMLS was initially used as a backup dictionary to identify ambiguous chemicals that did not match a synonym in PubChem. Ostensibly, the combined dictionaries would give greater coverage of the chemical namespace. In practice, however, the UMLS add little value to the chemical matching process and was relegated to a minor role; namely, it was used to manually identify ambiguous chemicals (see Chapter 4.2.1). PubChem is the primary means of resolving chemical synonymy and mapping chemicals to unique identifiers.

3.3 Creating a Database of Consumer Products

3.3.1 Verifying that the Data Owner Allows Scraping

Many online retailers sell consumer products. Drugstore.com (since acquired by Walgreens) had an extensive inventory and also provided ingredient lists on product webpages. Fortunately, at the time of writing, Walgreens' terms of use and robot exclusion protocol (https://www.walgreens.com/robots.txt) still allow web scraping under terms similar to those of Drugstore.com. Most online retailers allow web scraping so that shopping and advertising bots can direct customers to their sites. After confirming that data collection was consistent with the retailer's terms of use and that robotic scraping was not prohibited, consumer product data were collected from Drugstore.com. Their terms of use state:

> *"You agree that your use of robots, spiders, crawlers, wanderers, Web agents and other such automated processes on the Site will be Standard for Robot Exclusion (SRE)-compliant robots ('robots') and when connecting to the Site, prior to downloading or indexing any pages on the Site, such robots will immediately visit http://www.drugstore.com/robots.txt ('the robots.txt file'). You understand that the robots.txt file is the only means by which robots are authorized to access the Site. … You agree not to reproduce, duplicate, copy, sell, resell or exploit for any commercial purposes, any portion of the Site…"*

Scraping is allowed as long as robots comply with the rules in their `robots.txt` file and scraped data are not redistributed or used for commercial purposes. The `robots.txt` file provides a sitemap to help robot scrapers navigate the site, a list of disallowed branches where

scrapers should not go, and a minimum crawl delay (in seconds) to avoid overwhelming the server with HTTP requests:

```
Sitemap: http://www.drugstore.com/Sitemaps/0/default.xml
User-agent: *
Disallow: /cart.asp
Disallow: /list.asp
Disallow: /onorder.asp
Disallow: /checkout/
Disallow: /user/
Disallow: /products/email_product.asp
Disallow: /products/writereview.asp
Disallow: /la/account/
Disallow: /la/order/
Disallow: /templates/HIPAA/info.asp
Disallow: /affiliate/content.asp
Disallow: /shoppingbag.asp
Disallow: /checkout/default.asp
Disallow: /popups/largerphoto/default.asp
Disallow: /pricing.asp
Disallow: /LookAheadSuggestions.aspx
Disallow: /templates/stdplist/default.asp
Disallow: /templates/stdcat/default.asp
Disallow: /templates/evgrndept/default.asp
Disallow: /templates/events/circular.asp
Disallow: /4213/edh
User-agent: adidxbot
Crawl-Delay: 1
```

3.3.2 Scraping the Online Retailer

The sitemap in the `robots.txt` file gives the URLs to webpages on the site. The web scraper uses these URLs to request product pages from the online retailer. The consumer product retrieval process is shown schematically in Figure 4.

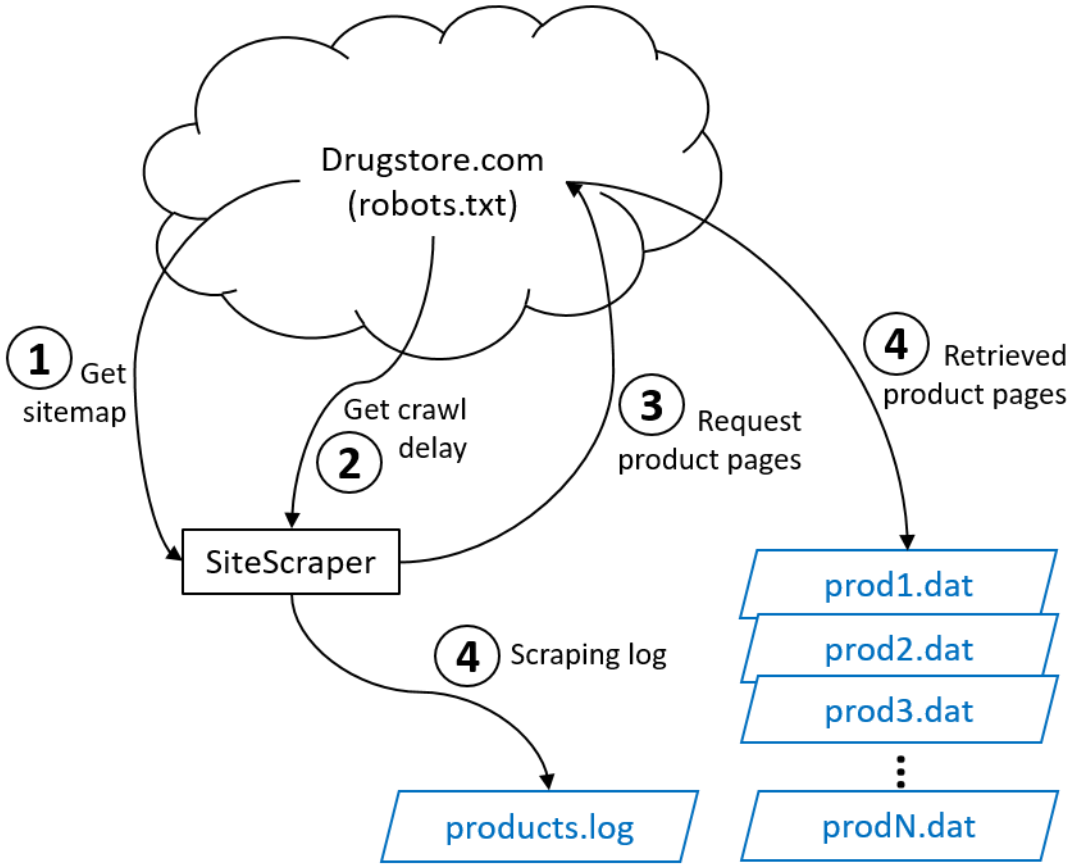


Figure 4 Web scraping process

After verifying that the target website allows the required data to be scraped, the program retrieves the sitemap and crawl delay, then sends page requests sequentially to the website. The raw HTML product files are saved to disk along with a log of successful page requests. Rectangles indicate computational processes, parallelograms indicate files, and arrows indicate data flow. Blue parallelograms indicate data used in subsequent stages.

The web scraping program, `SiteScraper`, consists of approximately 130 lines of Java and uses the Apache HttpClient (version 3.1) to request product pages. The target website is specified by setting the `robotsURL` variable inside the `SiteScraper.java` source file. In this case, it is set to `http://www.drugstore.com/robots.txt`, though this URL is now defunct since Walgreens acquired Drugstore.com. The program is compiled and run as follows:

```
Compile: javac -d  bin -sourcepath src \
               -cp bin:httpcomponents-client-4.3.3/lib/commons-codec-1.6.jar: \
                       httpcomponents-client-4.3.3/lib/commons-logging-1.1.3.jar: \
                       lib/commons-httpclient-3.1.jar src/SiteScraper.java

Run: java -cp bin:httpcomponents-client-4.3.3/lib/commons-codec-1.6.jar: \
                  httpcomponents-client-4.3.3/lib/commons-logging-1.1.3.jar: \
                  lib/commons-httpclient-3.1.jar SiteScraper
```

`SiteScraper` should run on any system where the Java virtual machine is supported. Note that HttpClient is no longer supported but its functionality has been incorporated into Apache HttpComponents. New development should use this package or some other supported HTTP request client (e.g., Jsoup, BeautifulSoup, cURL).

Scraping took several days given the size of the Drugstore.com product catalog and the required crawl delay. The robot exclusion protocol specified a one-second crawl delay but this was doubled to put less strain on their servers. Scraping is a network-limited rather than a compute- or memory-limited process, so a powerful server with specialized hardware is not necessary. A reliable network connection and sufficient disk space are more important because scraping tens of thousands of product pages transfers and consumes many gigabytes of data.

3.3.3 Extracting the Required Information from the Raw HTML

Brand and product names, ingredient list, and product category are needed for this analysis. This information is available on most Drugstore.com product pages (Figure 5) and can be extracted from the raw HTML retrieved by the robot scraper. This is done by finding tags that consistently mark the desired information across a given retail site. For example, the "TblProdForkIngredients" tag indicates the location of the product ingredient list in Drugstore.com product pages (Figure 6).

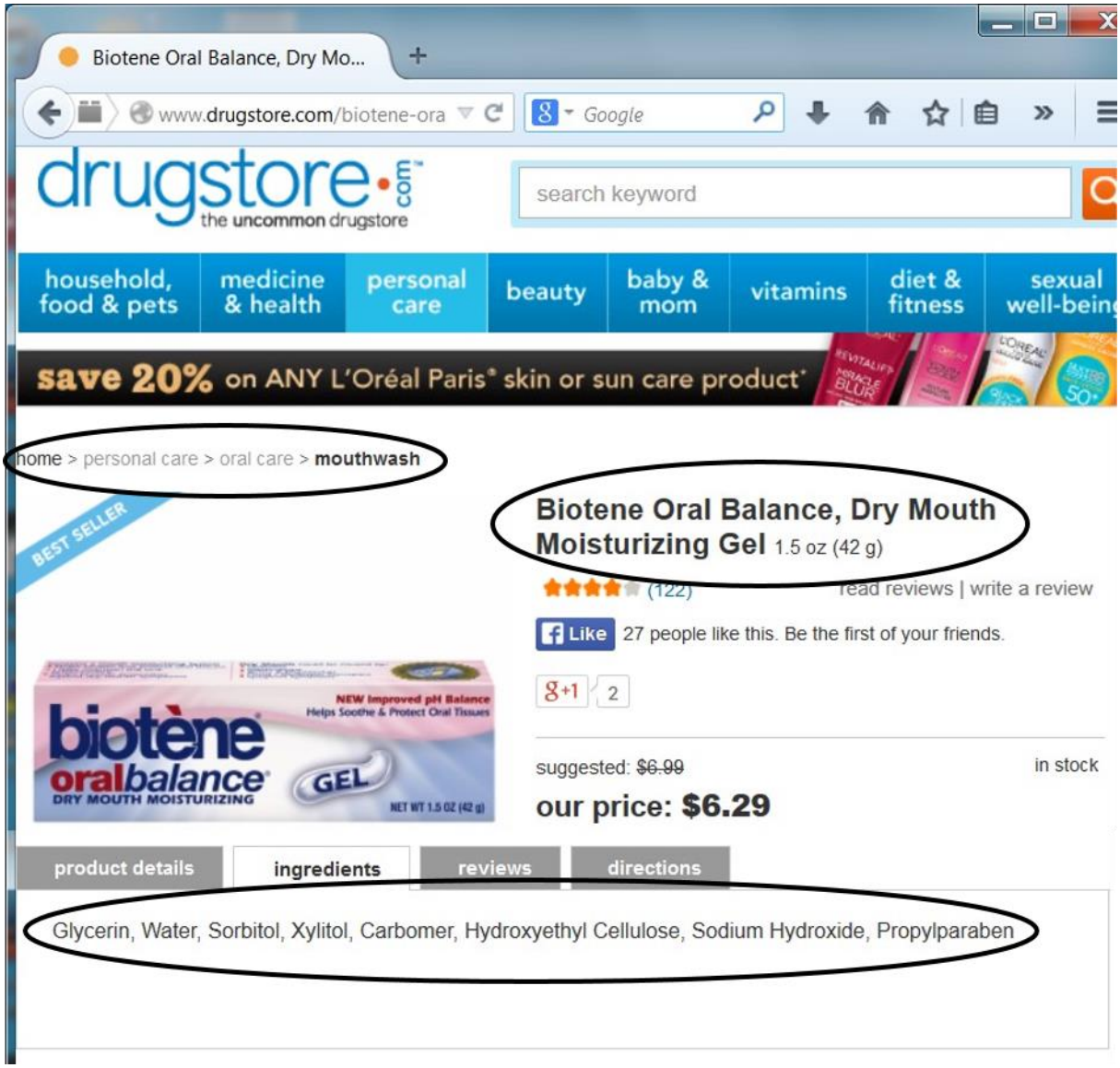


Figure 5 Example product webpage from Drugstore.com

This example shows the critical information (circled) that must be extracted: brand name, product name, retail hierarchy, and ingredient list.

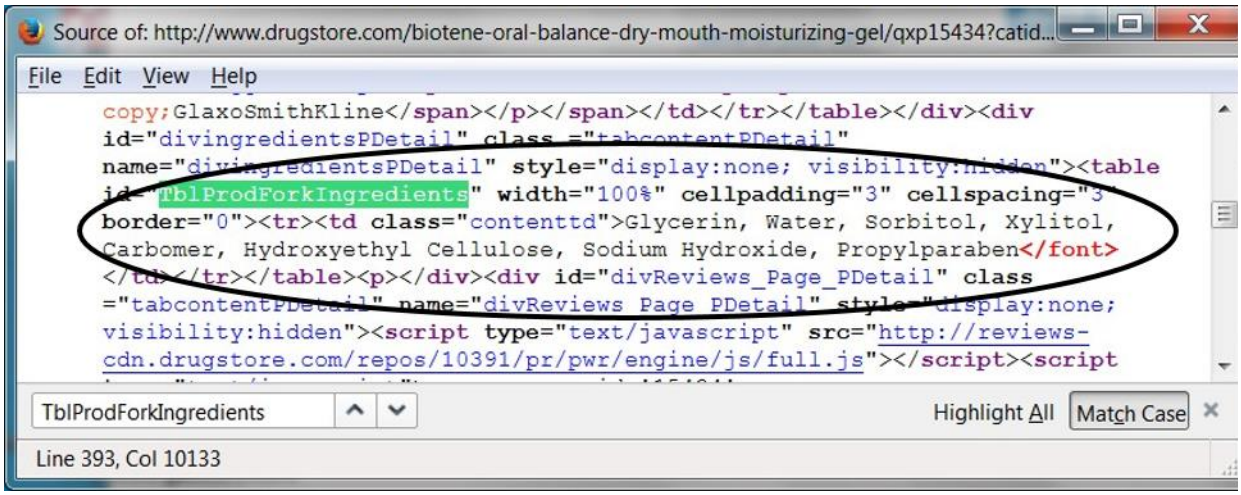


Figure 6 Example of raw HTML for Drugstore.com product webpages

This example shows the tag (TblProdForkIngredients) used to find and extract the product ingredient list from the raw HTML.

The first occurrence of the "s.prop5" and "<title>" tags indicate the brand and product names, respectively, and the "home<" tag indicates the retail hierarchy for product categorization (e.g., home → personal care → oral care → mouthwash). Assigning product categories using the retail hierarchy is described below. These tags vary by retailer but once identified are consistent and reliable across a given retailer's product pages.

Frequent spot checks of random samples are used to refine each stage of data processing. Validation of brand and product names was performed by manual inspection of 100 randomly selected products to confirm that the necessary data was correctly extracted from the raw HTML. Accuracy was 100% (i.e., every brand and product name in the sample was correct).

Category assignments were similarly validated using a random sample of 100 products. Accuracy was high (96%). Of the four incorrectly categorized products, one was due to an error in the retail hierarchy; specifically, an eyeliner product was incorrectly placed in the lip liner branch of the sitemap. The rest were due to ambiguities in category mapping. For example, one of the incorrect assignments was a topical medication in a relatively sparse branch of the retail hierarchy: medicine & health → pain & fever relief → shop by active ingredient → natural ingredients. The most specific level of the retail hierarchy that maps to one of our product categories is "pain & fever relief" so it was used to make the assignment. In our categorization

scheme, “pain & fever relief” maps to oral medications because most products in this category are oral medications.

A combination of Python, regular expressions, grep, and the html2text utility were used to process the raw HTML product pages. Extracting the brand names, product names, and product categories was straightforward but extracting the ingredients required more finesse because there is no standard format for ingredient lists. Most product labels provide a simple comma-delimited list of ingredients. However, some lists contain non-ingredient text, active concentrations, and parenthetical information that may or may not be useful, e.g.:

```
active ingredients: avobenzone – 2 % (sunscreen),
homosalate (15%), octisalate (5%) (sunscreen), oxybenzone –
4 % (sunscreen) inactive ingredients: alcohol denat,
acrylates, octylacrylamide, glycerin, aloe barbadensis leaf
extract, tocopherol (vitamin e), cocos nucifera oil
(coconut), mineral oil, fragrance
```

Simply processing this string as a comma-delimited list will result in noisy ingredient names that are more difficult to match to chemicals. However, patterns in such strings inform a multistep text processing algorithm that yields a clean list of ingredients for most product label formats.

Step 1: Remove “`active ingredients:␣`” (the ␣ symbol denotes a single space) and replace “`␣inactive ingredients:␣`” with a comma:

```
avobenzone – 2 % (sunscreen), homosalate (15%), octisalate
(5%) (sunscreen), oxybenzone – 4 % (sunscreen),alcohol
denat, acrylates, octylacrylamide, glycerin, aloe
barbadensis leaf extract, tocopherol (vitamin e), cocos
nucifera oil (coconut), mineral oil, fragrance
```

Step 2: Parse the comma-delimited list with the following regular expression to get a preliminary list of ingredient strings:

```
Regex = ((?:(?:[^,]+? +?\(.+?\))+(?:[^,]+)?)|(?:[^,]+))
```

```
avobenzone - 2 % (sunscreen)
homosalate (15%)
octisalate (5%) (sunscreen)
oxybenzone - 4 % (sunscreen)
alcohol denat
acrylates
octylacrylamide
glycerin
aloe barbadensis leaf extract
tocopherol (vitamin e)
cocos nucifera oil (coconut)
mineral oil
fragrance
```

Step 3: Product labels often contain extraneous text like “usp”, “denat” or “denatured”, “certified organic”, “contains less than”, etc. so a list of the most common non-ingredient phrases was compiled. Such text is removed in this step.

Step 4: Extract active concentrations from the ingredient strings using the regular expression below. Note that active concentrations are specified in percentages, milligrams, or units. Active concentrations are not used in the present analysis but they are retained for future use.

```
Regex = ([0-9\.|\,0-9]*\s?)(%|mg|units)
```

```
avobenzone - (sunscreen)
homosalate
octisalate (sunscreen)
oxybenzone - (sunscreen)
alcohol
acrylates
octylacrylamide
glycerin
aloe barbadensis leaf extract
tocopherol (vitamin e)
cocos nucifera oil (coconut)
mineral oil
fragrance
```

Step 5: Extract parenthetical text using the regular expression below. Parenthetical text often contains information that can help identify chemical ingredients so it is retained in the `cpdb_product_ingredient_paren.csv` file shown in Figure 7 and Table 5. Any leftover trailing punctuation is also removed in this step to yield a final, clean list of ingredient names.

```
Regex = \(([^)]+)\)
```

```
avobenzone
homosalate
octisalate
oxybenzone
alcohol
acrylates
octylacrylamide
glycerin
aloe barbadensis leaf
extract
tocopherol
cocos nucifera oil
mineral oil
fragrance
```

The ingredient string processing algorithm was validated by randomly selecting 100 products for manual inspection. Parsed ingredient lists were compared to the raw ingredient strings to confirm that ingredient names and accompanying parenthetical text are correctly extracted. Of the 1,587 ingredients in this sample, 1,547 (97%) were correctly extracted. Of the 40 incorrectly extracted ingredients, 24 were slash-delimited polymers, fatty acids, or mixtures (e.g.: styrene/acrylates copolymer, acrylates/c10 30 alkyl acrylate crosspolymer, cetyl peg/ppg-10/1 dimethicone, caprylic/capric triglyceride, pvm/ma copolymer). The ingredient string processing algorithm was not modified to handle these types of ingredients because they are not the focus of the present analysis and because it is unclear how they should be parsed. Missing commas in the ingredient list caused the remaining 16 incorrectly parsed ingredients.

The output of the web scraping step of the workflow consists of the raw HTML product pages and a log of successful page requests, as shown in Figure 4. The scraping log is a pipe-delimited text file. Each record consists of two fields: a unique product identification number assigned by `SiteScraper` and the URL of the product that was scraped, e.g.:

```
1|http://www.drugstore.com/johnsons-baby-moisture-wash/qxp163501
2|http://www.drugstore.com/johnsons-baby-take-along-pack/qxp10796
```

The raw data is stored in a subdirectory called `raw_product_pages`. It contains an HTML file (i.e., `prod1.dat`, `prod2.dat`, etc.) for each record in `products.log`. The critical data for this research (brand names, product names, ingredient lists, and product categories) is extracted from the raw HTML using a multistep data extraction and cleaning workflow (Figure 7). The components of this workflow are described in Table 5.

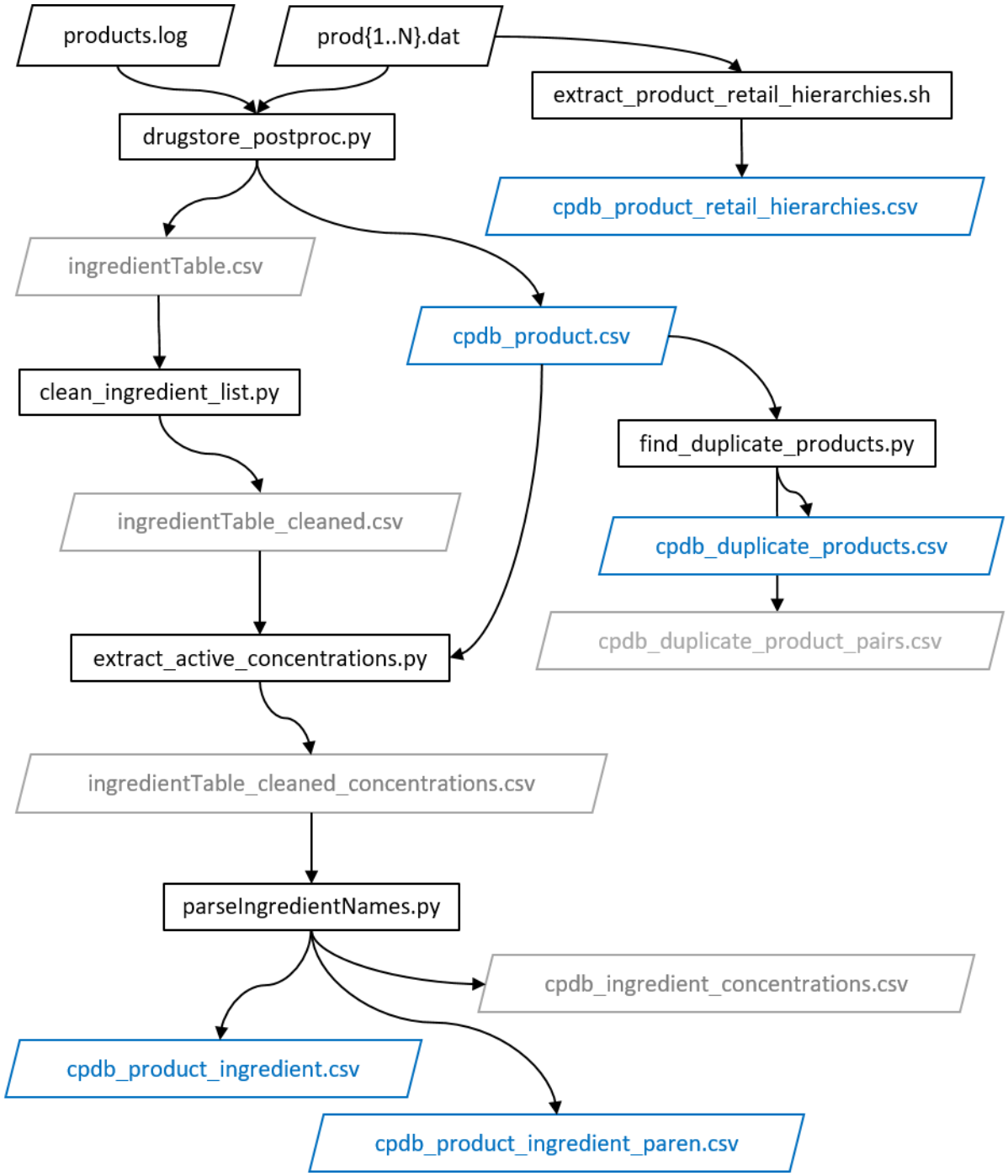


Figure 7 Extracting and cleaning data from the raw consumer product HTML files

Rectangles indicate computational processes, parallelograms indicate files, and arrows indicate data flow and dependencies. Blue parallelograms indicate data used in subsequent stages. Gray parallelograms indicate intermediate files, validation data, or data that are not used in subsequent stages.

Table 5 Extracting and cleaning data from the raw consumer product HTML files

Computational Processes	
File Name	**Description**
drugstore_postproc.py	This workhorse program extracts the brand name, product name, and raw ingredient string from each `prodN.dat` file. The ingredient string is parsed into individual ingredients. Run this program in the `raw_product_pages` subdirectory.
`Command: python drugstore_postproc.py < products.log`	
clean_ingredient_list.py	This program removes non-ingredient strings (e.g., extraneous text like "may contain" or "certified organic" that is not part of an ingredient name) from the ingredient list.
`Command: python clean_ingredient_list.py < ingredientTable.csv \` `> ingredientTable_cleaned.csv`	
extract_active_concentrations.py	Some products, particularly medications, specify concentrations for active ingredients. This program parses this information, even though it is not used in the present research.

`Command: python extract_active_concentrations.py \`
`cpdb_product.csv \`
`ingredientTable_cleaned.csv \`
`> ingredientTable_cleaned_concentrations.csv`	

parseIngredientNames.py	This program performs the final cleaning of the ingredient list by separating the basic ingredient name, any parenthetical descriptors that can help identify the ingredient, and active concentrations (when specified).
`Command: python parseIngredintNames.py < ingredientTable_cleaned_concentrations.csv`	
find_duplicate_products.py	For various reasons, there may be duplicates in the product list. Duplicates must be identified to avoid overcounting product categories and ingredients. This program finds duplicate products using the algorithm described in Chapter 4. The Dice coefficient and Levenshtein ratio thresholds are set in this program.

Table 5 (cont.)

`Command: python find_duplicate_products.py < cpdb_product.csv \`
`1> cpdb_duplicate_product_pairs.csv \`
`2> cpdb_duplicate_products.csv`	

extract_product_retail_hierarchies.sh	This script extracts and parses the retail hierarchies from the raw HTML product pages. It invokes two small Python programs (`extract_product_hierarchy_step1.py` and `extract_product_hierarchy_step2.py`) and the Python `html2text` utility. Run this script in the `raw_product_pages` subdirectory.
`Command: bash extract_product_retail_hierarchies.sh`	
Data	
File Name	**Description**
cpdb_product.csv	Each record in this file contains the product ID, brand name, product name, raw ingredient list, retrieval date, and URL in pipe-delimited fields.
cpdb_duplicate_product_pairs.csv	This pipe-delimited file contains pairs of duplicate products. Each record specifies the IDs of the two products, the Dice coefficient of the two product names, and the Levenshtein ratio of the two ingredient lists.
cpdb_duplicate_products.csv	This file is simply the set of duplicate products, one product ID per record.
cpdb_product_retail_hierarchies.csv	Each pipe-delimited record in this file contains a product ID, the level of the retail hierarchy, and a description for that branch and level of the retail hierarchy.
cpdb_product_ingredient.csv	This file contains the product ID, ingredient number (i.e., position in the ingredient list), and ingredient name in pipe-delimited fields.

Table 5 (cont.)

cpdb_product_ingredient_paren.csv	This pipe-delimited file has the same fields as cpdb_product_ingredient.csv but contains only the parenthetical text associated with some ingredients. This information helps to identify the ingredient when the base name is insufficient.
ingredientTable.csv, ingredientTable_cleaned.csv, ingredientTable_cleaned_concentrations.csv	Intermediate files containing the product ingredients in progressive stages of data cleaning.
cpdb_ingredient_concentrations.csv	Each record in this file contains the product ID, ingredient number, ingredient concentration, and unit of measurement in pipe-delimited fields. This data is not currently used in this research.

Computational processes and files in the workflow to extract and clean data from the raw consumer product HTML files. Backslashes indicate command-line continuation.

3.4 Incorporating Consumer Product Usage Patterns and Usage Modes

Similar to Cowan-Ellsberry and Robison (2009) and Comiskey et al. (2015), Kantar Worldpanel (http://www.kantarworldpanel.com/global/Consumer-Panels) profiles are used to determine dominant product categories and product combinations based on actual consumer usage, and the Scientific Committee on Consumer Safety (SCCS, 2015) scaling factors are used to estimate likely chemical retention (i.e., absorption into the body). However, the present work includes a large list of specific chemicals rather than just parabens or generic fragrance. When combined with the CPDB, it is possible to prioritize for risk assessment the specific chemical ingredients and ingredient combinations based on near-field, cumulative chemical exposure (and retention) from consumer products.

The Kantar Worldpanel monitors consumer behavior worldwide, primarily for marketing purposes. Participants track their daily product usage for one week and submit detailed diaries. The Kantar subset used in the present study consists of the weighted average daily usage patterns (by product category) of 11,000 American consumers (age 13+) collected between October 2014 and September 2015. The structure of this dataset is shown in Figure 8. Usage

profiles from actual consumers show which product categories and combinations are used most often. When combined with the CPDB, usage profiles will show the ingredients and ingredient combinations that contribute most to near-field chemical exposure from consumer products. The Kantar dataset allows the determination of chemical combinations based on combined product usage, whereas Gabb and Blake (2016a) could only determine combinations on a per-product basis.

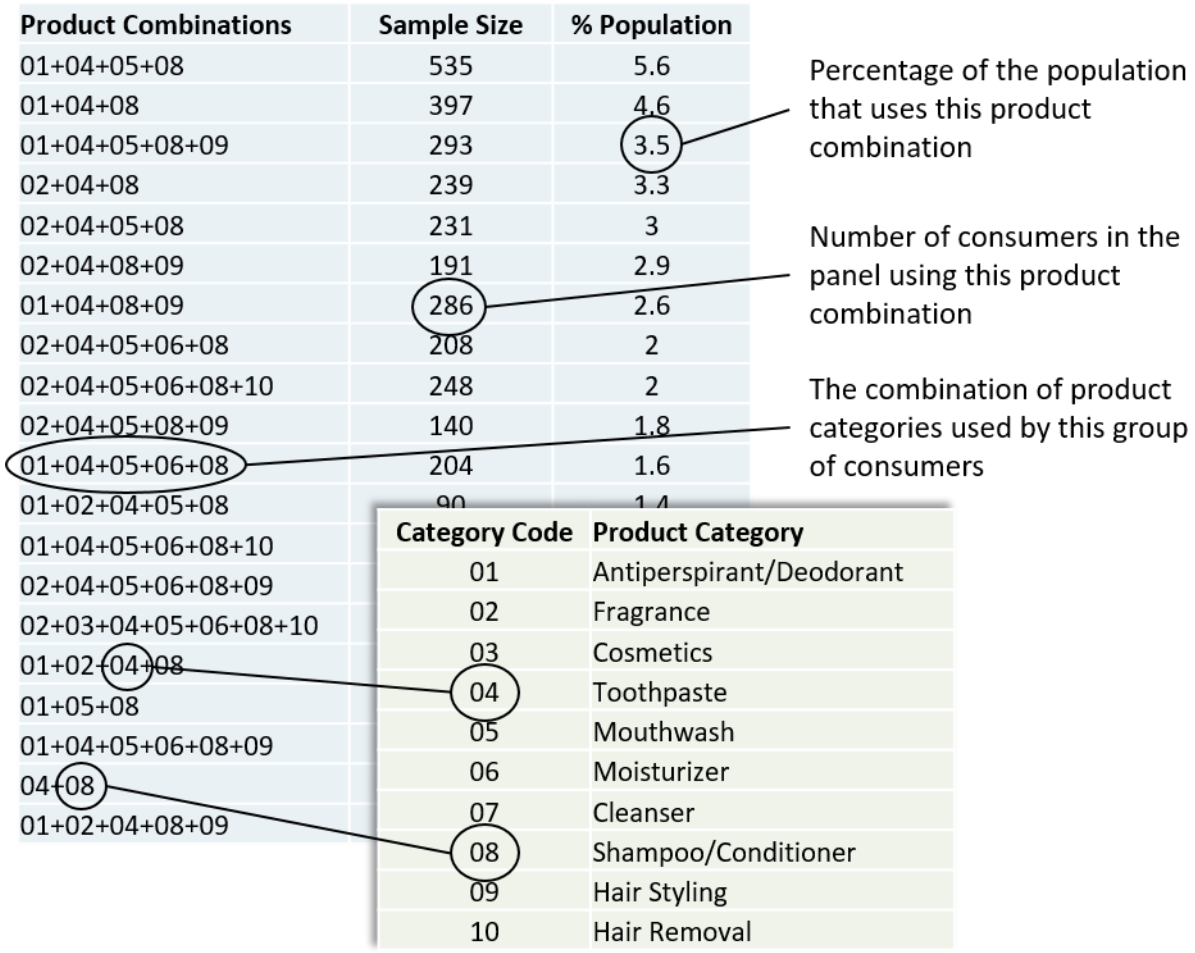


Figure 8 Structure of the Kantar Worldpanel dataset used in this research (not actual data)

The European Statistical Population Model (ESPM) is a stochastic exposure model for cosmetics and personal care products (Hall et al., 2007; McNamara et al., 2007; Hall et al., 2011). This model results from a large, longitudinal study of European consumers, in which 80,000 households and 14,413 individuals from five countries provided their product usage information for 12 product categories. Loretz et al. (2005, 2006, 2008) created an exposure

model similar to the ESPM in terms of product categories and intent, but theirs was compiled from a significantly smaller population sample (a few hundred American consumers) using more detailed product usage diaries (e.g., the hair length of shampoo users, the skin type of lotion users). Both of these models provide daily estimated exposure for each product category for male and/or female consumers, but neither model considers the specific chemicals or even chemical classes that occur in consumer products.

Chapter 4: Data Cleaning

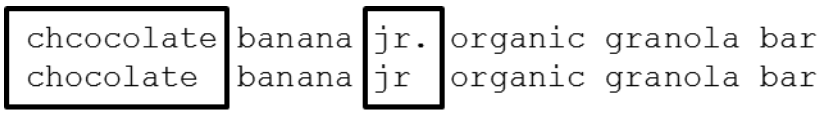
Gathering and integrating the necessary datasets is only the first step of any informatics workflow. The next step is cleaning that data. In practice, datasets are typically messy. Among other things, they have holes where data is missing or layout and type inconsistencies that disrupt data extraction. All data is sociotechnical, so it has a human component. It would be a mistake to assume that born-digital datasets are somehow less messy because even machine-generated data has a human foundation – the programmer, the lab technician, etc. – and is therefore subject to human error. Data scraped from online sites is no exception. Retail websites like Drugstore.com are developed by humans, and product webpages retrieve data from databases populated by human data entry clerks.

Accurate counting and categorization of consumer products are essential to achieve the goal of prioritizing the target chemicals for CRA. Chapter 4.1 describes how duplicates are removed from the consumer products dataset to avoid overcounting. The process by which product categories are assigned, and the accuracy of this process, are also described, along with a final breakdown of the consumer product sample used in this analysis. To detect and tabulate the target chemicals in the product sample, it is essential that both the target chemicals and the product ingredients are accurately identified. Chapter 4.2 describes how chemical names are mapped to unique identifiers. Ambiguities in the chemical namespace complicate this task, so the processes by which synonymy and homonymy are handled are also described.

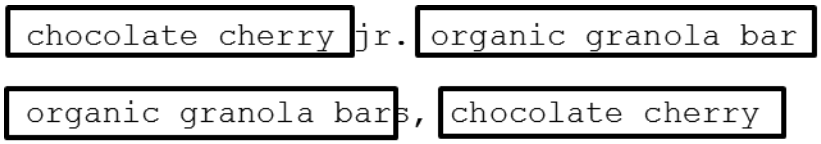
4.1 Cleaning the Consumer Product Database

4.1.1 Removing Duplicate Products from the Database

Duplicate products can appear in the database for several reasons. The same product can appear in different branches of a retail sitemap. The same product may be sold in different sizes. In the future, as more retail sites are scraped and added to the database, product inventories may overlap, leading to duplicate entries. Pruning duplicates is necessary to get accurate counts of products and ingredients, but identifying duplicate products is not always as straightforward as matching product names under the same brand because typographical errors and differences in punctuation can mask duplicates, e.g.:



Unfortunately, digital text contains typographical errors just like printed text. If these two products have identical brands and ingredient lists, they are likely the same product scraped from different locations. Alternative word orders in product names can also mask duplicate products, e.g.:



It is harder to identify these two products as duplicates because the words, word order, and punctuation are different. However, if they also have identical brands and ingredient lists, they are likely different representations of the same product. Applying a spelling checker to fix typographical errors, removing punctuation, and doing string matching on the product names will find many duplicate products but it will not find duplicates when the word order of the product names differ. Dice's coefficient (Dice, 1945) is a better way to compare product names in this case:

$$Dice\ coef. = \frac{2|S_1 \cap S_2|}{|S_1| + |S_2|}\ where\ S_N\ is\ the\ set\ of\ character\ bigrams\ in\ string\ N$$

If two product names have a high Dice coefficient, they are not necessarily the same product because formulations change. Their ingredient lists must still be compared. Labeling regulations dictate that ingredients be listed in descending order of predominance, so word order matters when comparing ingredient lists. Therefore, the Levenshtein ratio (Navarro, 2001) is a better way to measure ingredient list similarity. It is computed as follows:

$$Levenshtein\ ratio = \frac{length(S_1) + length(S_2) - ED(S_1, S_2)}{length(S_1) + length(S_2)}\ where\ ED\ is\ the\ edit\ distance\ between\ strings\ S_1\ and\ S_2$$

Edit distance (Navarro, 2001) was computed using the edit_distance function in the Natural Language Toolkit (https://www.nltk.org). The algorithm to find duplicate products is shown in Figure 9. Manual analysis of the 5,426 brands in the product sample found only ten (0.2%) incorrectly specified brand names (e.g., "slim-fast" instead of the correct "slimfast"). Only 125 products are associated with incorrect brands. Even if all of them are duplicates, their effect on the tabulation of the target chemicals will be minimal. Therefore, exact matching was used to compare brand names rather than more computationally expensive partial string matching techniques (i.e., Dice coefficient and Levenshtein ratio).

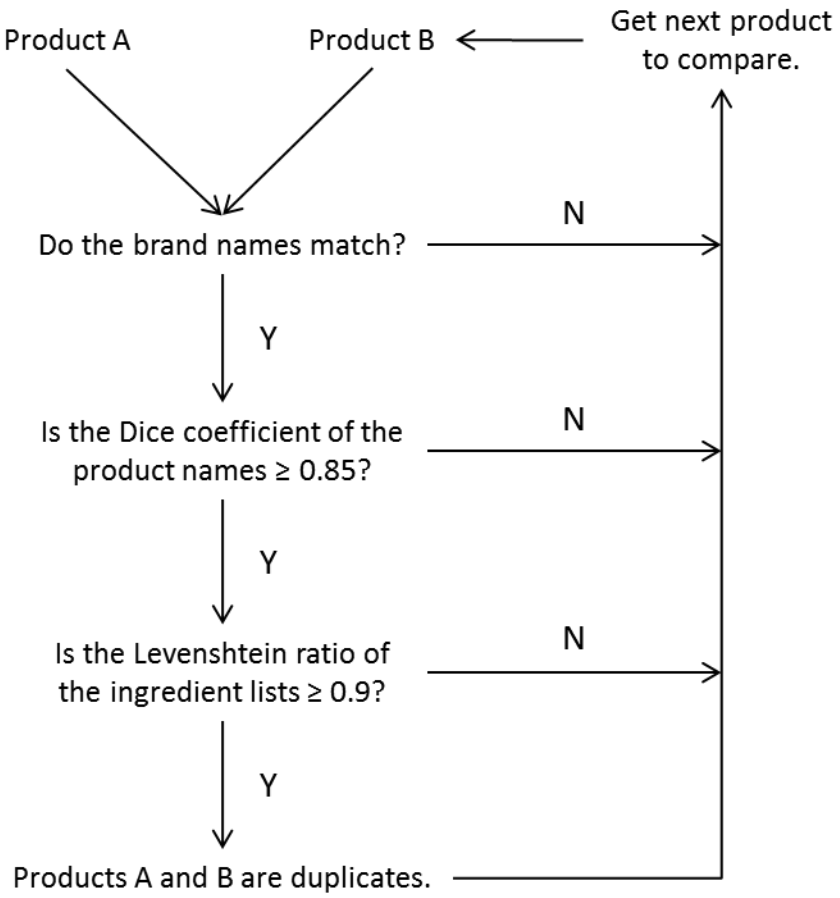


Figure 9 Algorithm used to find duplicate products

The algorithm was tuned and validated using a manually curated sample. A random sample is unlikely to contain duplicate products so ten brands with ten products each were selected and manually analyzed for duplicates. The sample contained 89 distinct and 11 duplicate products. The selected duplicates contained typographical errors and product name

variations. Dice coefficient and Levenshtein ratio thresholds of 0.85 and 0.9, respectively, gave the best results, correctly identifying 9 out of 11 duplicates with no false positives.

4.1.2 Assigning Product Categories

Consumer products have various usage modes. Some are left on after application, while some are rinsed off. Some are used on the hair, while some are used in the mouth. Some are applied to the skin, while some are applied to the eyes or mucosa. A product's category usually says something about its usage mode, which in turn says something about the likely retention of chemical ingredients. Therefore, accurately assigning product categories is critical for more than just cataloging the sample. Gabb and Blake (2016a) used an approach similar to Goldsmith et al. (2014) to annotate product categories. Product pages on retail sites typically include the product's location in the retailer's hierarchy. For example, toothpaste might be in the home→personal care→oral care→toothpaste branch of the retail hierarchy. This information is included to help customers navigate the site and shop more efficiently. It is used here to categorize products because retail categories are objective, and retailers have a vested interest in making sure they are correct.

Products were assigned to their respective categories as follows. First, a dictionary is created from all terms/phrases (e.g., personal care, oral care, toothpaste) in the Drugstore.com retail hierarchies scraped in April 2014, May 2015, March 2016, and September 2016. This resulted in a dictionary of 1,685 terms/phrases. The dictionary is sorted in descending order of frequency to show the most important terms and phrases (i.e., those that affect the largest number of consumer products), e.g.:

```
body wash|3411
lotions|3342
hair color|3230
liquid foundation|2551
salon styling products|1573
lipstick|1234
shaving cream|433
smoothing & frizz control|256
...
```

Annotation involves reviewing each term/phrase and deciding whether it maps unambiguously to a particular product category. Those that unambiguously mapped to one of the desired product categories were kept. For example, the phrase “regular mascara” maps to the “Make Up” category; “all women’s fragrance” maps to the “Fragrance” category; “bar soap” maps to “Body Washing/Soap/Cleansers;” and so forth. Ambiguous phrases (e.g., “Personal care”) were discarded. This left 223 terms/phrases that define product categories. Assigning a category to a product is usually straightforward, but some products can exist in more than one category (e.g., products labeled as “shampoo and conditioner” or “shampoo and body wash”). If a product could be assigned to more than one category, the most specific level of its retail hierarchy (e.g., toothpaste in the home→personal care→oral care→toothpaste branch) is used to make the final assignment. Products that did not map to one of the categories were assigned “Other.” Once these decisions are made, frequencies in the dictionary are replaced by categories, e.g.:

```
body wash|Cleanser
lotions|Moisturizer
hair color|Styling
liquid foundation|Makeup
salon styling products|Styling
lipstick|Makeup
shaving cream|Shaving
smoothing & frizz control|Styling
...
```

Figure 10 shows the workflow to assign product categories. The components of this workflow are described in Table 6.

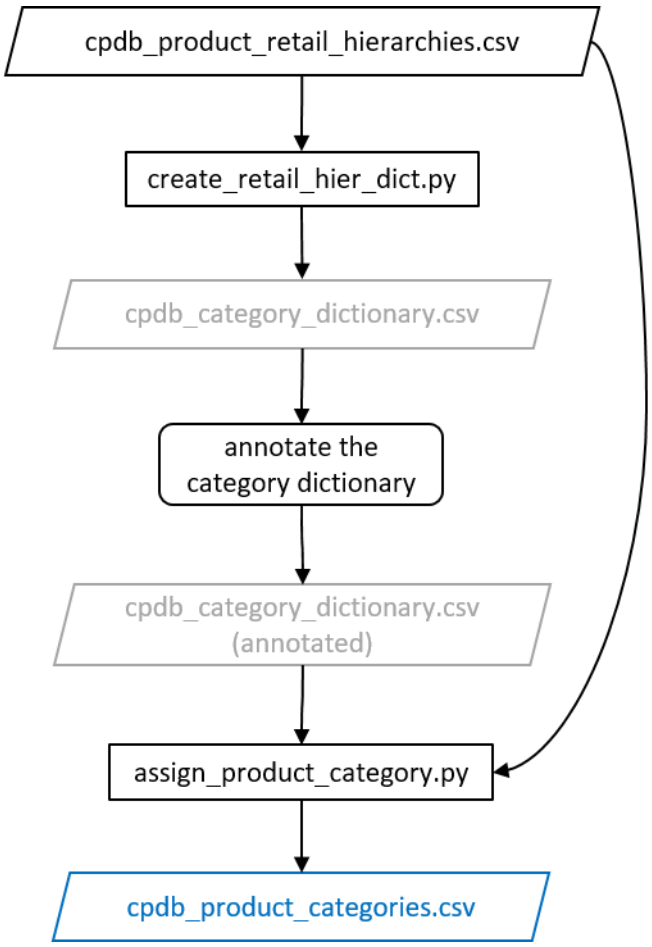


Figure 10 Assigning categories to the consumer products

Rectangles indicate computational processes, rounded rectangles indicate manual processes, parallelograms indicate files, and arrows indicate data flow and dependencies. Blue parallelograms indicate data used in subsequent stages. Gray parallelograms indicate intermediate or validation data that are not used in subsequent stages.

Table 6 Assigning product categories

Computational Processes	
File Name	**Description**
create_retail_hier_dict.py	This program outputs a dictionary of terms/phrases from the retail hierarchy sorted in descending order of frequency.
`Command: python create_retail_hier_dict.py < cpdb_product_retail_hierarchies.csv \` `> cpdb_category_dictionary.csv`	
assign_product_category.py	This program uses the annotated category dictionary to assign categories to products based on their retail hierarchies. In cases where a product can be mapped to more than one category, the higher-level (i.e., more specific) term/phrases from the retail hierarchy is used to make the assignment.

`Command: python extract_active_concentrations.py \`
`< cpdb_product_retail_hierarchies.csv \`
`> cpdb_product_categories`	

Data	
File Name	**Description**
cpdb_product_retail_hierarchies.csv	See Table 5.
cpdb_category_dictionary.csv	Initially, each record in this pipe-delimited file contains a term/phrase from the retail hierarchy and its frequency. After manual annotation, each record should contain the term/phrase and the product category to which it belongs.
cpdb_product_categories.csv	Each pipe-delimited record in this file contains a product ID, the level of the retail hierarchy used to make the category assignment, and the assigned product category.

Computational processes and files in the workflow to assign product categories. Backslashes indicate command-line continuation.

This approach worked well. Validation was done by selecting a random sample of 500 products and manually checking their assigned categories. All but 19 were correct (96% accuracy, the same accuracy reported in Gabb and Blake, 2016a). Of the incorrectly categorized

products, one was due to an error in the retail hierarchy; specifically, an eyeliner product was incorrectly placed in the lip liner branch of the hierarchy. The rest were due to ambiguities in category mapping. For example, one of the incorrect assignments was a topical medication in a relatively sparse branch of the retail hierarchy: medicine & health → pain & fever relief → shop by active ingredient → natural ingredients. The most specific level of the retail hierarchy that maps to one of our product categories is "pain & fever relief" so it was used to make the assignment. In this categorization scheme, however, "pain & fever relief" maps to oral medications because most products in this category are oral medications. The 19 incorrect assignments showed where improvements could be made mapping retail hierarchy terms/phrases to product categories. These improvements were made and the categorization process was repeated. Validation of another random sample of 500 products showed 15 incorrect assignments (97% accuracy). Further refinements did not improve accuracy.

4.1.3 Tabulating the Product Sets

The categories and sample sizes in the original CPDB (Gabb and Blake, 2016a) are shown in Table 7. As much as possible, products were mapped to one of the categories used by Dodson et al. (2012). Five of their categories (cat litter, pillow protectors, vinyl shower curtains, car interior cleaners, and car air fresheners) were excluded because the CPDB did not contain any representative products. Their household cleaning categories (i.e., surface, floor, tub and tile, and glass cleaners and scrubbing powder) were also combined into a single category (i.e., cleaner) because the sample sizes of the specific categories are small relative to the other household categories in Table 7. Combining them into a single category helps to balance sample sizes within the broad household category. Finally, several categories (mostly under medication and diet) were added to accommodate products in the CPDB that were not tested by Dodson et al. (2012).

Table 7 Product sample size for the original CPDB (Gabb and Blake, 2016a)

Broad Category	Specific Category	Number of Products	Percentage Containing One or More DODSON Chemicals	Number of DODSON Chemicals in Category
household	air fresheners	197	15.3	4
	cleaner	108	5.5	3
	diapers	72	2.1	1
	dishwashing	121	14.2	7
	laundry	273	3.3	6
	pesticide	158	10.0	7
	pet supplies	612	2.1	3
	other	395	5.7	9
personal cleaning	bar soap	620	6.3	11
	body wash	1075	33.4	18
	facial cleanser	622	57.5	19
	hand sanitizer	44	11.3	4
	liquid soap	289	29.7	9
	other	501	44.0	10
personal care	body oil & body spray	231	28.2	12
	deodorant & antiperspirant	518	12.3	13
	feminine hygiene	237	23.1	8
	lotion & moisturizer	2467	66.5	19
	sexual health	333	23.6	7
	shaving & hair removal	480	34.3	16
	sunscreen	503	71.8	14
	other	1094	51.6	19
oral care	mouthwash	154	24.7	3
	toothpaste	332	12.8	9
hair care	conditioner	1363	58.4	20
	hair color	256	48.9	10
	hair styling	1479	63.3	18
	shampoo	1338	43.9	19
	other	53	48.3	11
cosmetics	bronzers & tanners	189	69.3	13
	eye makeup	1688	66.8	15
	foundation	1657	72.3	14
	fragrance & perfume	505	51.4	12
	lip makeup	1606	42.3	13
	manicure & pedicure	1792	14.9	22
	other	243	62.6	13
medication	oral medication	1957	7.3	13
	topical medication	772	25.8	14
	other	360	10.0	6
diet	food	3324	0.8	2
	supplements	4291	1.2	6
	tea	610	3.1	1
	vitamins	3583	0.9	4
other	other	473	14.9	12

Product categories, sample sizes, and the percentage of products in each category that contain at least one of the DODSON chemicals, and the number of target chemicals appearing in each product category.

The database contains 41,277 products that have at least one ingredient listed on the product label. Exact duplicates (the same brand and product name scraped from different locations) and partial duplicates (different sizes of the same product) were pruned to avoid inflating ingredient counts, as described previously. The final database used in Gabb and Blake (2016a) contained 38,975 distinct products (from 8,099 brand names) with 32,231 distinct ingredient names, of which 7,486 mapped to a CID and/or CUI after resolving synonymous names (e.g., water, eau, agua, distilled water, purified water, etc.). This is much larger than the 8,921 products with 1,797 unique chemicals found in a database constructed by scraping Material Safety Data Sheets (MSDS) (Goldsmith et al., 2014). In contrast to MSDS, for which products are only required to list those ingredients known to be hazardous, the database used here includes all ingredients listed on a product label.

The original CPDB retrieved from Drugstore.com in April 2014 (83,730 products) was updated with additional "scrapes" from May 2015 (73,577 products), March 2016 (64,372 products), and September 2016 (44,345 products). (Note that the shrinking inventory was a likely harbinger of the absorption of Drugstore.com into the larger Walgreens retail site.) After removal of duplicate products and those that do not typically provide an ingredient list (e.g., vinyl shower curtains, plastic storage containers, toothbrushes, makeup brushes), the new CPDB contained 55,209 distinct products compared to 38,975 in the original sample. The present analyses used this updated database except where otherwise noted. The final breakdown of personal care products by category is shown in Table 8. These ten product categories coincide with those in the consumer product usage patterns purchased from Kantar.

Table 8 Breakdown of personal care products by category in the Kantar dataset

Product Category	Number of Products
Makeup	11,099
Face/body moisturizer	5,050
Hair shampoo/conditioner	3,846
Body wash/soap/cleanser	3,705
Hair styling	2,543
Fragrance	1,111
Antiperspirant/deodorant	842
Toothpaste	749
Shaving and hair removal (gel, foam, etc.)	619
Mouthwash	250
Other	24,011
Retail hierarchy not found	1,384

Personal care products were assigned to one of ten categories corresponding to those in the consumer usage patterns in the Kantar dataset.

4.2 Cleaning and Using the Chemical Dictionary

4.2.1 Mapping the Target Chemicals to Unique Identifiers

Accurately mapping the target chemicals and product ingredients to unique identifiers is critical to this research. Good techniques are available to recognize chemicals in free text, ranging from simple dictionary-based approaches (Hettne et al., 2009) to machine learning (Degtyarenko et al., 2008; Grego et al., 2012; Hawizy et al., 2011; Jessop et al., 2011; Klinger et al., 2008; Leaman et al., 2015). The tmChem tool (Leaman et al., 2015) represents the current state-of-the-art based on blind critical assessment (Krallinger et al., 2013, 2015a, 2015b). However, complex approaches to chemical entity recognition are unnecessary for the present work. Much of the complexity of these approaches lies in identifying the correct start and end points of chemical strings in free text. This is unnecessary for the target chemicals because the text strings (and sometimes even the CID) are already provided in their source lists.

Product ingredient lists are also more structured than free text. They are simply comma-delimited lists that can be parsed using ordinary text processing, as described in Chapter 3.3.3. In scientific articles, chemicals can appear as trivial names, systematic names, or abbreviations.

On product labels, the relative brevity of trivial names is advantageous, as the following example illustrates. Galaxolide is one of many trivial names for the fragrance chemical 4,6,6,7,8,8-hexamethyl-1,3,4,7-tetrahydrocyclopenta[g]isochromene (systematic name). Dictionary-based matching works well for trivial names but not as well for systematic names (Hettne et al., 2009; Jessop et al., 2011; Klinger et al., 2008), and parsing systematic names is another source of complexity in chemical entity recognition (Lowe et al., 2011). Fortunately, systematic names are rare in consumer product labels (Gabb and Blake, 2016a) so complex parsing based on chemical morphology is unnecessary.

Therefore, a dictionary-based, exact matching approach is used to map chemical names to unique identifiers. Chemical dictionaries are appropriate and effective because they often have dozens, sometimes hundreds, of synonyms. For example, PubChem contains approximately 67 million CIDs and 131 million synonyms at the time of writing. A trivial name appearing in a product ingredient list is likely to be among those synonyms. Exact matching is appropriate and effective because sophisticated partial string matching techniques (e.g., Dice's coefficient, edit distance, and Levenshtein ratio) (Dice, 1945; Navarro, 2001) are prone to false positives and false negatives when dealing with chemical names. For example, "vitamin a" and "vitamin e" are similar strings but different chemicals (false positive), whereas "dimethyl ether" and "methoxymethane" are dissimilar strings but the same chemical (false negative).

Figure 11 shows the steps to process PubChem Compound and use it to assign unique identifiers to the target chemicals. The components of this workflow are described in Table 9. Target chemicals that do not match a PubChem synonym must be manually identified and mapped to a PubChem CID, as described below.

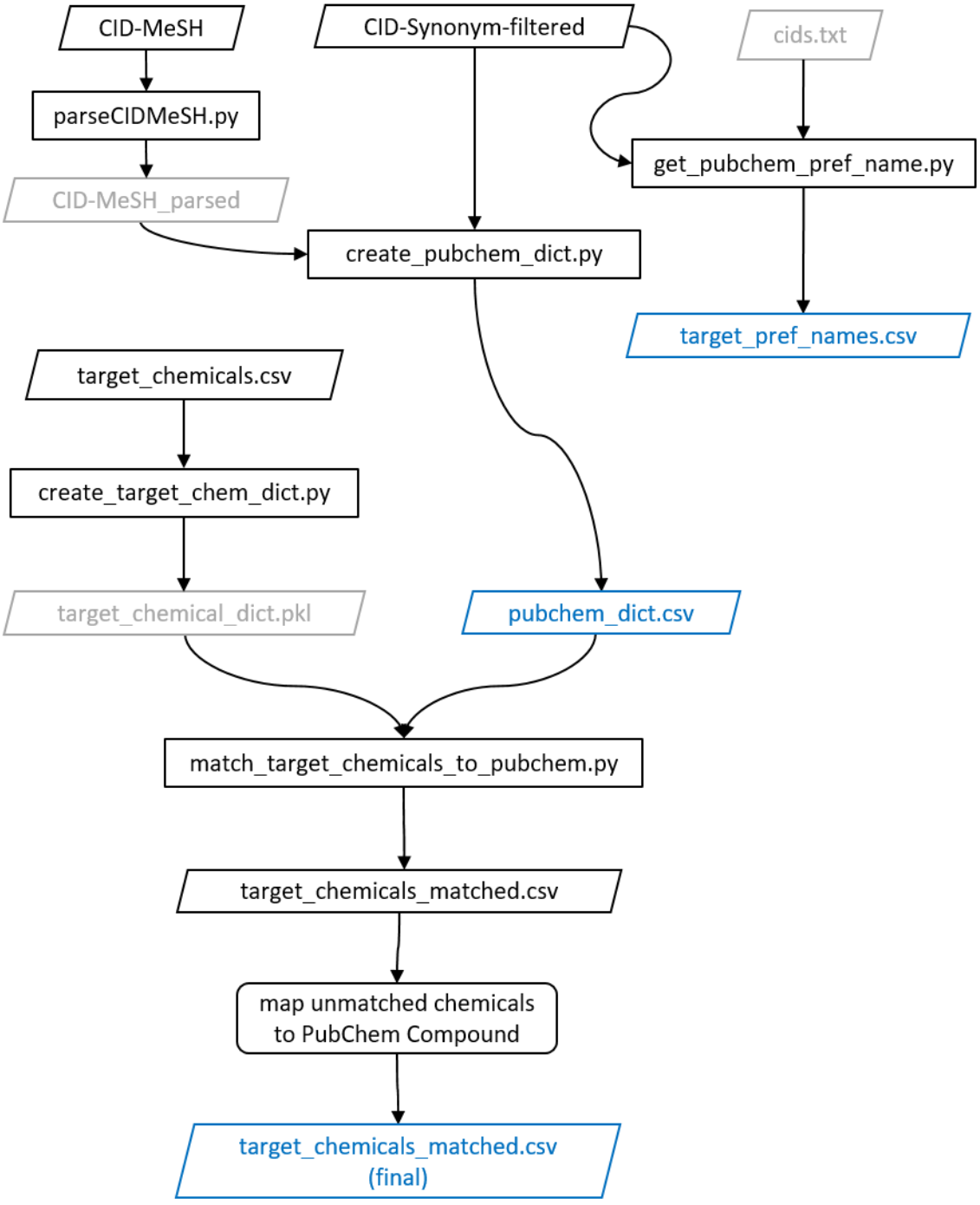


Figure 11 Mapping the target chemicals to PubChem CIDs

Rectangles indicate computational processes, rounded rectangles indicate manual processes, parallelograms indicate files, and arrows indicate data flow and dependencies. Blue parallelograms indicate data used in subsequent stages. Gray parallelograms indicate intermediate or validation data that are not used in subsequent stages.

Table 9 Mapping the target chemicals to PubChem CIDs

Computational Processes	
File Name	**Description**
get_pubchem_pref_name.py	This program simply extracts the first PubChem synonym, generally a preferred or commonly used name, for each target chemical.
`Command: python get_pubchem_pref_name.py > target_pref_names.csv`	
create_pubchem_dict.py, parseCIDMeSH.py	These workhorse programs perform all of the editing and preprocessing of PubChem synonyms (e.g., normalizing to lowercase, removing extraneous text, syntactic inversion) described in Chapter 4 to improve chemical name matching.

`Command: python parseCIDMeSH.py < CID-MeSH > CID-MeSH_parsed`
`cat CID-Synonym-filtered CID-MeSH_parsed |  \`
`python create_pubchem_dict.py > pubchem_dict.csv`	

create_target_chem_dict.py	This program performs transformations on the target chemical names similar to those performed on the PubChem synonyms by `create_pubchem_dict.py`.
`Command: python create_target_chem_dict.py < target_chemicals.csv`	
match_target_chemicals_to_pubchem.py	This program matches the target chemicals to PubChem to assign CIDs.
`Command: python match_target_chemicals_to_pubchem.py > target_chemicals_matched.csv`	
Data	
File Name	**Description**
CID-Synonym-filtered	This tab-delimited file contains the synonyms associated with each PubChem CID. This file is downloaded from PubChem Compound.
CID-MeSH, CID-MeSH_parsed	These tab-delimited files contain the MeSH terms associated with each PubChem CID. They are added to the final synonym dictionary (`pubchem_dict.csv`). This file is downloaded from PubChem Compound.

Table 9 (cont.)

pubchem_dict.csv	This tab-delimited, key:value store contains the CID (value) associated with each PubChem synonym (key).
target_chemicals.csv	See Table 2.
cids.txt	This file is simply a list of CIDs extracted from `target_chemicals.csv`.
target_chemical_dict.pkl	This Python pickle file contains the serial number and source list of each target chemical (key) and its set of synonyms extracted from the authoritative lists (value).
target_chemicals_matched.csv	This tab-delimited file contains the serial number and source list of the target chemical, its assigned CID, and its matching PubChem synonym.
target_pref_names.csv	This pipe-delimited file contains the CID and PubChem preferred name for each target chemical.

Computational processes and files in the workflow to map the target chemicals to PubChem CIDs. Backslashes indicate command-line continuation.

A random sample of 100 matched chemicals did not find any that were mapped to incorrect identifiers. Spot checks of unmapped chemicals suggested additional preprocessing steps to improve matching of chemical names to PubChem synonyms: removing trailing, bracketed text and ignoring unnecessary dashes and whitespace, as described in Chapter 3.2.

Of the 17,856 entries in the authoritative lists (TOX21: 9,011, HSDB: 5,731, CACOC: 2,444, EDCDB: 615, DODSON: 55), 16,408 matched a PubChem synonym; 95 were positively identified and manually mapped to a CID; 1,302 were excluded because the entry was a mixture rather than a distinct chemical compound, was a protein or nucleic acid, or was ambiguous and could not be positively identified [e.g., the entry “t-butylphenyl diphenyl phosphate” could map to “2-t-butylphenyl diphenyl phosphate” (CID: 158333) or “4-t-butylphenyl diphenyl phosphate” (CID: 70425)]; and 51 were positively identified but were not in PubChem (Table 10). Entries that did not match a synonym in PubChem were identified using ChemSpider

(http://www.chemspider.com), ChemIDplus (https://chem.nlm.nih.gov/chemidplus/; Tomasulo, 2002), SciFinder (https://scifinder.cas.org), and/or the UMLS. There is considerable overlap among the lists, so their union provided the final list of 11,964 distinct target chemicals for this research (Figure 12). The complete chemical lists and the PubChem CID mappings are provided in the Supplemental Material (Target Chemicals and Target Chemicals Overlap).

Table 10 Final breakdown of the authoritative lists of target chemicals

List	Entries	Entries Mapped to a PubChem CID		Unmatched/Excluded Entries			
		Automatic	Manual	Mixture	Protein/DNA	Unidentified	Not in PubChem
TOX21	9,011	8,801	34	131	3	39	3
HSDB	5,731	5,518	17	118	48	2	28
CACOC	2,444	1,419	44	885	10	66	20
EDCDB	615	615					
DODSON	55		55				

In general, most entries matched a synonym in PubChem and did not require manual analysis. Entries that did not match a PubChem synonym were either positively identified and manually mapped to a CID or excluded with some justification. Only 107 out of 17,856 (0.6%) entries were ambiguous and could not be positively identified. Only 51 chemicals (0.3%) were missing from PubChem. Chemicals in the EDCDB are already mapped to PubChem CIDs. The DODSON chemicals were manually mapped to CIDs in Gabb and Blake (2016a).

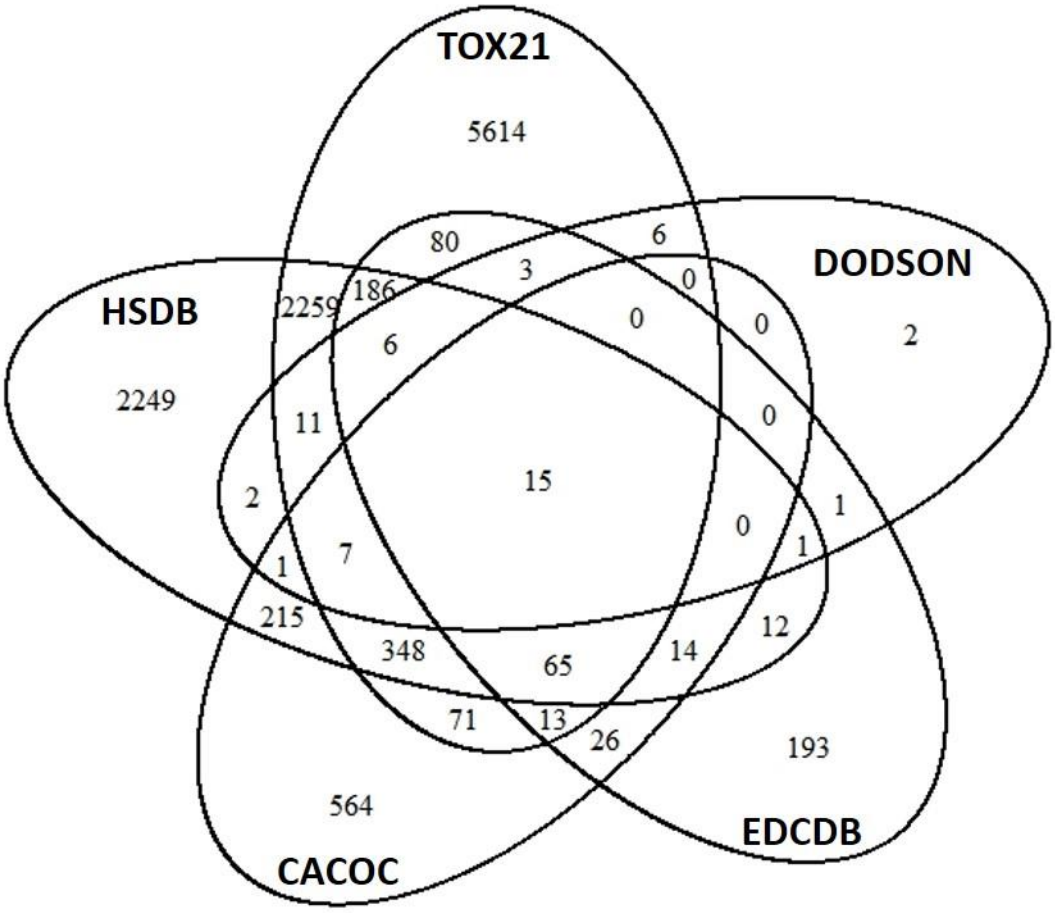


Figure 12 Overlap among the authoritative lists of potentially harmful chemicals

The numbers in the diagram indicate the number of chemicals in each set. For example, there are 15 chemicals in the TOX21 ∩ HSDB ∩ CACOC ∩ EDCDB ∩ DODSON set while there are 2,259 chemicals common to just TOX21 and HSDB. Notice that each list contains chemicals that do not appear in any other list. The complete table of overlapping chemicals is provided in the Supplemental Material (Target Chemical Overlap).

4.2.2 Mapping Product Ingredients to Unique Identifiers

The ingredients parsed from product labels were mapped to unique chemical identifiers using the same dictionary-based, exact matching approach that was used for the target chemicals. Figure 13 shows the workflow to assign unique identifiers to the product ingredients. The components of this workflow are described in Table 11. Unlike the authoritative lists of target chemicals, parsing the ingredient lists can be noisy because they are not all simple comma-delimited lists. Many lists contain parenthetical and/or non-ingredient text that must be extracted or otherwise handled in order to expose the actual ingredient names. Therefore, all ingredient strings that matched a target chemical in PubChem were

manually examined for correctness. Of the 2,117 strings that mapped to one of the 11,964 target chemicals, only 112 were found to be incorrect (95% accuracy). The incorrectly mapped ingredient strings were filtered from subsequent tabulations to avoid inflating the ingredient counts or detecting chemicals that are not really present in consumer products. Table 12 shows the breakdown of the incorrectly mapped ingredient strings. The complete analysis is included in the Supplemental Material (Ingredient Validation).

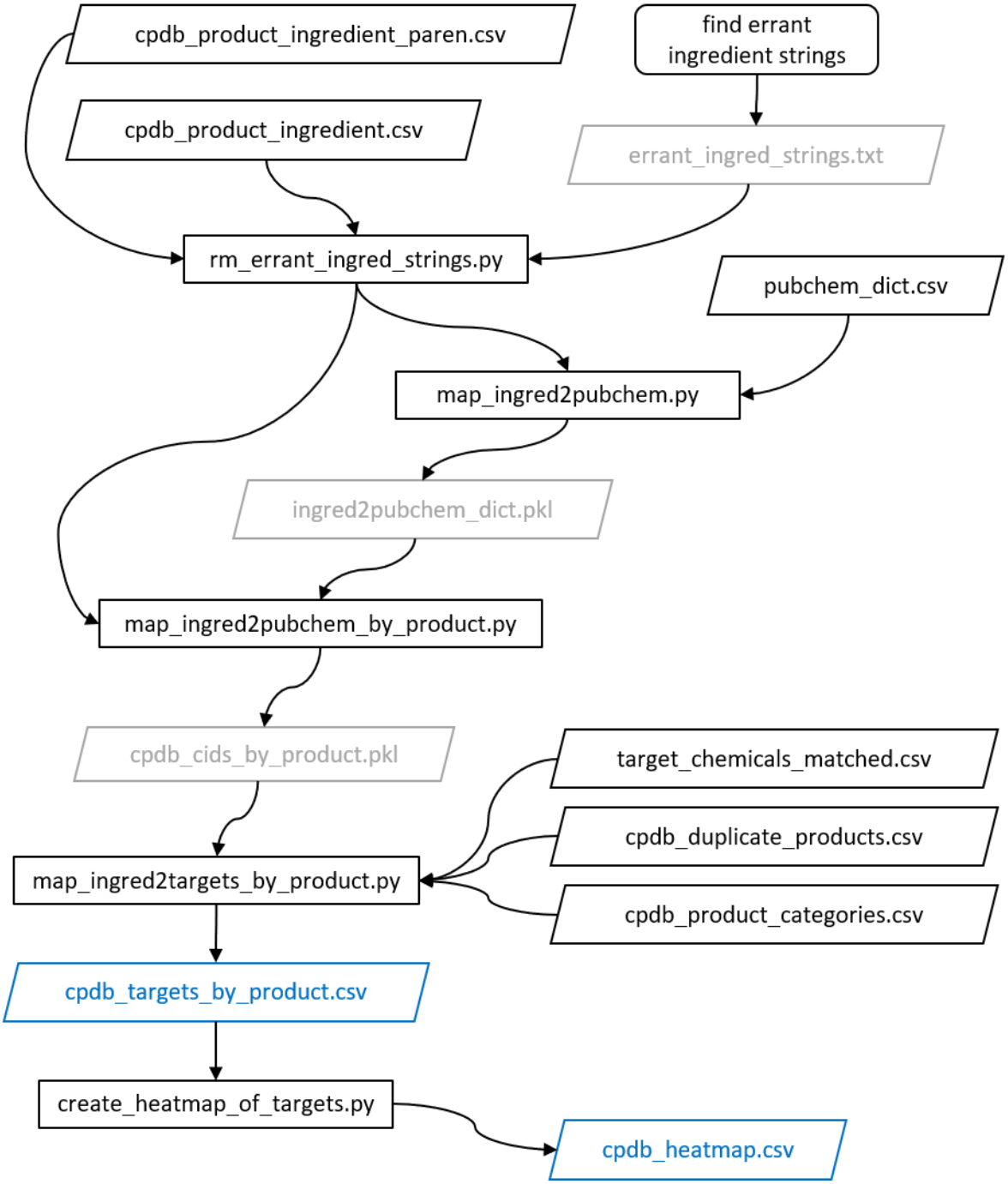


Figure 13 Generating the matrix of target chemical proportions by product category

Rectangles indicate computational processes, rounded rectangles indicate manual processes, parallelograms indicate files, and arrows indicate data flow and dependencies. Blue parallelograms indicate data used in subsequent stages. Gray parallelograms indicate intermediate or validation data that are not used in subsequent stages.

Table 11 Tabulating target chemical proportions by product category

Computational Processes	
File Name	**Description**
rm_errant_ingred_strings.py	This simple script filters ingredient strings that are known to map incorrectly to PubChem synonyms. It is typically used in combination with `map_ingred2pubchem.py` and `map_ingred2pubchem_by_product.py`.
`Command: cat cpdb_product_ingredient.csv cpdb_product_ingredient_paren.csv | \` `python rm_errant_ingred_strings.py`	
map_ingred2pubchem.py	This program matches product ingredient strings to PubChem synonyms, and hence to CIDs.

`Command: cat cpdb_product_ingredient.csv cpdb_product_ingredient_paren.csv | \`
`python rm_errant_ingred_strings.py | \`
`python map_ingred2pubchem.py > ingred2pubchem_dict.csv`	

map_ingred2pubchem_by_product.py	This program maps the ingredients in each consumer product to their PubChem CIDs.

`Command: cat cpdb_product_ingredient.csv cpdb_product_ingredient_paren.csv | \`
`python rm_errant_ingred_strings.py | \`
`python map_ingred2pubchem_by_product.py > cpdb_cids_by_product.csv`	

map_ingred2targets_by_product.py	This program identifies the target chemicals in each consumer product in the database.
`Command: python map_ingred2targets_by_product.py > cpdb_targets_by_product.csv`	
create_heatmap_of_targets.py	This program tabulates the frequency and computes the proportion of each target chemical in each product category.
`Command: python create_heatmap_of_targets.py < cpdb_targets_by_product.csv \` `> cpbd_heatmap.csv`	
Data	
File Name	**Description**
cpdb_product_ingredient.csv, cpdb_product_ingredient_paren.csv, cpdb_duplicate_products.csv	See Table 5.

Table 11 (cont.)

cpdb_product_categories.csv	See Table 6.
errant_ingred_strings.txt	This file contains strings parsed from the raw ingredient lists that map incorrectly to PubChem synonyms. They are found during manual validation of ingredient-to-PubChem mapping (Table 12). There is one errant string per record.
pubchem_dict.csv	See Table 9.
ingred2pubchem_dict.pkl	This intermediate file contains the strings parsed from the product ingredient lists and the PubChem CID to which they map.
cid_by_product_dict.pkl	Each record in this intermediate key:value store contains the set of target chemicals CIDs (value) for a product ID (key).
cpdb_targets_by_product.csv	Each pipe-delimited record in this file contains a product ID, a product category, and a list of target chemical CIDs for that product.
target_chemicals_matched.csv	See Table 9.
cpdb_heatmap.csv	This critical file contains the data for every target chemical that is detected in the CPDB. Each pipe-delimited record contains a CID, a product category, the number of products in that category where the chemical is detected, and the proportion of products in this category that contain the chemical.

Computational processes and files in the workflow to tabulate target chemical proportions by product category. Backslashes indicate command-line continuation.

Table 12 Manual validation of ingredient string matching to PubChem

Error Type	Number	Examples
Valid match to PubChem synonym but the ingredient refers to a different substance	48	The ingredient string “lime” refers to the fruit but maps to calcium oxide. Interestingly, nine of these incorrect matches are due to street names for controlled substances (i.e., illegal drugs). For example, “chocolate chips” is a street name for LSD so this ingredient is incorrectly mapped to lysergic acid.
Valid match to PubChem synonym but the ingredient string denotes a function rather than a specific chemical	42	The ingredient strings “anti-dandruff” and “anti-acne” are synonyms for salicylic acid but they are too vague to confirm the presence of this chemical in the product.
Incorrect mapping due to parsing artifact	13	Forward slashes in ingredient strings (e.g., caprylic/capric triglyceride) sometimes cause problems during parsing, as noted in Gabb and Blake (2016a). Fortunately, this is rare.
Valid match to PubChem synonym but the ingredient string is too vague to confirm a specific chemical	9	The ingredient strings “fat” and “fatty acid” map to specific chemicals in PubChem but really denote chemical classes.

Most of the ingredient strings that matched a chemical in PubChem were valid (2,005 out of 2,117, 95% accuracy). The 112 ingredient strings that incorrectly matched a chemical in PubChem were removed prior to the analysis.

In addition to checking for false positive matches, an analysis of unmatched ingredient strings was done to assess the degree to which false negatives can dampen the signal from the target chemicals. In this case, a false negative is an ingredient string representing a valid chemical name that fails to map to a PubChem CID. Parsing the product files results in approximately 2.4 million ingredient strings (the cpdb_product_ingredient.csv and cpdb_product_ingredient_paren.csv files in Figure 7 and Table 5). Of these, approximately 39,000 are unique, and about 2,000 match a PubChem synonym. It is infeasible to manually analyze the 37,000 unmatched ingredient strings, so the 500 most frequently occurring strings were examined (Table 13). The complete analysis is included in the Supplemental Material (Unmatched Ingredient Analysis).

Table 13 Manual analysis of unmatched ingredient strings

Reason Not Matched	Number	Example
Not a distinct chemical	329	Fragrance/flavor mixture, plant extract, animal product, polymer of variable length
Synonym not in PubChem	112	Mica, ci 77491, red 7 lake, dimethiconol, fd&c blue #1
Not an ingredient, stray text from ingredient string	30	Certified organic, all natural, may contain, plant based
Parsing artifact	14	Ingredients containing slashes or commas
Chemical not in PubChem	12	Cyclopentasiloxane, caprylyl glycol
Misspelled ingredient name	3	Butylphenyl methylproprional, vitamin a palminate, ethylhexyl glycerin

Most of the unmatched ingredient strings are mixtures rather than distinct chemicals. However, many unmatched strings represent valid chemicals that are not in PubChem.

Most of the unmatched ingredient strings are mixtures rather than distinct chemicals, so they are not expected to match an entry in PubChem. However, some of these ingredients, particularly fragrance and flavor mixtures, could be masking chemicals of interest. Also, the 124 missing synonyms and chemicals indicate that the PubChem, like most dictionaries, is incomplete. (Such limitations are discussed further in Chapter 7.2.) Of the 112 synonyms that were not found in PubChem, 44 are false negatives for a target chemical, resulting in some degree of signal loss. To put this in perspective, 1,147 of the target chemicals are detected in the consumer product sample. Only 37 of the missing synonyms have since been added to the latest version of PubChem. Of the 12 missing chemicals, seven have since been added to PubChem. Fortunately, none of the missing chemicals are among the targets, so they do not cause signal loss.

Colors and dyes account for 48 of the 112 missing synonyms. Eight result in signal loss for a target chemical. Colorants are a persistent problem for a number of reasons. First, many are mixtures of ground up minerals rather than distinct chemicals, so they are not included in PubChem. Second, many appear in ingredient labels as color indices (e.g., ci 77491) or FDA Food, Drug, and Cosmetic designations (e.g., fd&c blue #1, red 7 lake). PubChem has sparse and inconsistent coverage of these colorant synonyms. Third, the strings for these colorants are

messy in product labels. Their spacing and punctuation are inconsistent. For example, fd&c blue #1 can appear as fd&c blue 1, fdc blue 1, f d & c blue 1, blue #1, or simply blue 1. Similarly, ci 77491 can appear as c.i. 77491, ci77491, or simply 77491. This makes parsing difficult because there are no consistent patterns on which to build regular expressions for automatic text processing. The parser errs on the side of caution because over-processing the ingredient strings tends to introduce false positives. It is better to have a slightly dampened signal for a handful of target chemicals than to detect chemicals that are not really present in consumer products.

Most of the parsing artifacts are due to ingredients strings containing slashes or commas. As noted previously in Chapter 3.3.3, slashes typically occur in polymer or fatty acid names (e.g.: styrene/acrylates copolymer, acrylates/c10 30 alkyl acrylate crosspolymer, cetyl peg/ppg-10/1 dimethicone, caprylic/capric triglyceride, pvm/ma copolymer), which denote polymer mixtures rather than distinct chemicals. The remaining parsing artifacts were due to occasional systematic names, which contain commas that disrupt parsing of the comma-delimited lists. Fortunately, this is rare because trivial rather than systematic names are used for most chemical ingredients, as noted in Chapter 4.2.1. Only one of the parsing artifacts resulted in signal loss for a target chemical, so changing the parsing scheme is not warranted.

The three misspelled ingredient names look correct but are each off by one letter: vitamin a palminate instead of vitamin a palmitate, ethylhexl glycerin instead of ethylhexyl glycerin, and butylphenyl methylproprional instead of butylphenyl methylpropional. Only the latter caused signal loss. The other two chemicals are not targets. Implementing a chemical spelling checker may be possible, but nontrivial. The resulting signal loss from misspelled chemical names was not large enough to merit the effort.

4.2.3 Resolving Chemical Synonymy

Different consumer products often use different names for the same chemical ingredient. This creates a natural messiness in the consumer product labels. Synonymy arises from the normal uncontrolled growth of language; in this case, the language describing chemical entities where trivial names represent the “convenient general language” of everyday chemistry and systematic names represent the “legal language” (Tate, 1967). Put another way,

trivial names are simplified, common, or traditional chemical names that are not derived from a formal nomenclature, while systematic nomenclatures attempt to unambiguously convey both the chemical entity and its chemical makeup (Leigh, 2012). Chemicals can be listed on a product label using a systematic or trivial name. For example, methyl paraben is the trivial name of the common preservative chemical methyl 4-hydroxybenzoate (systematic name). Ambiguity in chemical nomenclature is a challenging problem that must be addressed in order to accurately determine the composition of consumer products.

The scientific sublanguage of chemistry evolves just like any other language, so synonymy and homonymy (covered separately in Chapter 4.2.4) are quite common in chemical nomenclature. The compound, methoxymethane, illustrates how synonymy arises in chemical names. This chemical is a product of wood distillation, so it was once commonly known as wood ether (i.e., an ethereal, or volatile, compound derived from woody substances). As the field matured and it became possible to determine chemical formula and structure, it was discovered that wood ether consisted of two methyl (i.e., $-CH_3$) groups connected by a bridging oxygen (i.e., CH_3-O-CH_3). Consequently, wood became synonymous with methyl (e.g., wood alcohol was once commonly used to refer to methanol, or methyl alcohol). Similarly, many volatile compounds contain a bridging oxygen so today, someone versed in organic chemistry nomenclature knows that the formula R-O-R' (where R and R' represent hydrocarbon groups) denotes an ether group. From this basic fact, the chemist also knows that:

1. Neither R nor R' can be a hydrogen because that would change the ether to a hydroxyl group (i.e., R-OH).
2. If R and R' are both hydrogen, the molecule (water) is inorganic and no longer subject to the rules of organic chemistry nomenclature.
3. Methyl ether and dimethyl ether must refer to the same chemical.
4. Monomethyl ether (i.e., one methyl group and one ether group) and trimethyl ether (three methyl groups and one ether group) are chemically impossible.

These pseudo-grammatical rules and transformations are instinctive to someone fluent in the chemistry sublanguage (Harris and Mattick, 1988). Wood ether, methyl ether, dimethyl ether,

methoxymethane, etc. are all valid names for the chemical CH_3-O-CH_3, so the choice of synonym becomes a matter of personal preference or context. For example, wood ether could be used for historical purposes or simply generational stubbornness. Methyl/dimethyl ether could be used when trivial names are appropriate (e.g., consumer product labels). Methoxymethane would be used when systematic names are required (e.g., chemistry journals).

While on the subject of language, it is worth mentioning in passing that the previous definition of chemical synonymy (i.e., different names for the same chemical) is not strictly correct. Synonymy, as used throughout this text, is actually coreference. Coreference is the relationship between language and the physical world. Synonymy describes terms that are related, and have the same sense, within a language. Chemical names refer to physical entities. Therefore, it is more correct to say that different names for the same chemical are coreferent rather than synonymous. This is a gross oversimplification, but a detailed philosophical and psycholinguistic discussion of sense and reference is outside the scope of this dissertation. A thorough literature review is similarly out of scope, but interested readers are referred to two classic works on the subject (Frege, 1892; Kripke, 1980). For practical reasons, the more familiar term, synonymy, is used instead of coreference.

Just over half (31 out of 55) of the DODSON chemicals appear in the original CPDB (Gabb and Blake, 2016a) (Table 1). Of these, 19 appear under more than one name. Therefore, synonymy must be taken into account in order to get an accurate count of products containing a particular ingredient. For example, bucinal is a fairly common synthetic fragrance, but searching ingredient lists for bucinal will miss all 539 products containing this chemical. Searching for its synonym, lilial (71 products), will still miss most of the products containing this chemical because it is more commonly listed as butylphenyl methylpropional (468 products). It is not intuitively obvious, even to a chemist, that bucinal, lilial, and butylphenyl methylpropional are synonyms. A lay consumer is even less likely to recognize chemical synonyms. Such is the case with many of the chemicals listed in Table 1, e.g.: octinoxate, benzophenone-3, decamethylcyclopentasiloxane, methyl salicylate, limonene, and 4-tert-octylphenol monoethoxylate. Methyl salicylate and limonene further illustrate the gap

between chemical names and ingredient labels. Though the chemical names are used most often, marketing factors may motivate the use of natural sounding names such as wintergreen oil or sweet birch oil instead of the chemical equivalent methyl salicylate. The Supplemental Material (Ingredient Validation) captures the degree of synonymy in product ingredient lists for all of the target chemicals. For example, benzyl alcohol (CID: 244), a target chemical from both TOX21 and HSDB, appears on product labels under the following synonyms: benzyl alcohol, benzoyl alcohol, phenylcarbinol, and hydroxymethylbenzene. Fortunately, as noted in Chapter 3.2, PubChem contains numerous synonyms for the chemicals in the database: approximately 150 million synonyms for 39 million CIDs. This large chemical dictionary combined with the mapping scheme described in Chapter 4.2.2 virtually ensure that synonymous ingredient names are assigned the same CID.

4.2.4 Accounting for Chemical Homonymy

Chemical synonymy, as defined previously, occurs when different names refer to the same chemical (e.g., vitamin e and tocopherol). Chemical homonymy occurs when the same name can refer to different chemicals [e.g., the generic name terpineol can refer to various stereoisomers or salts of the parent compound, 2-(4-methylcyclohex-3-en-1-yl)propan-2-ol]. The degeneracy of 2D molecular descriptors (i.e., different compounds sharing the same descriptor) is a known problem in chemistry (Faulon et al., 2005; Randic, 1984). Similarly, shared synonyms among the various salts and stereoisomers of a compound can lead to homonymy among PubChem CIDs (Figure 14). Thus, a chemical name can refer to more than one CID. However, this also means that when searching for a particular chemical among tens of thousands of consumer product ingredient lists, all the PubChem synonyms associated with that chemical plus the synonyms associated with its homonymic CIDs are available for possible matching (Table 14).

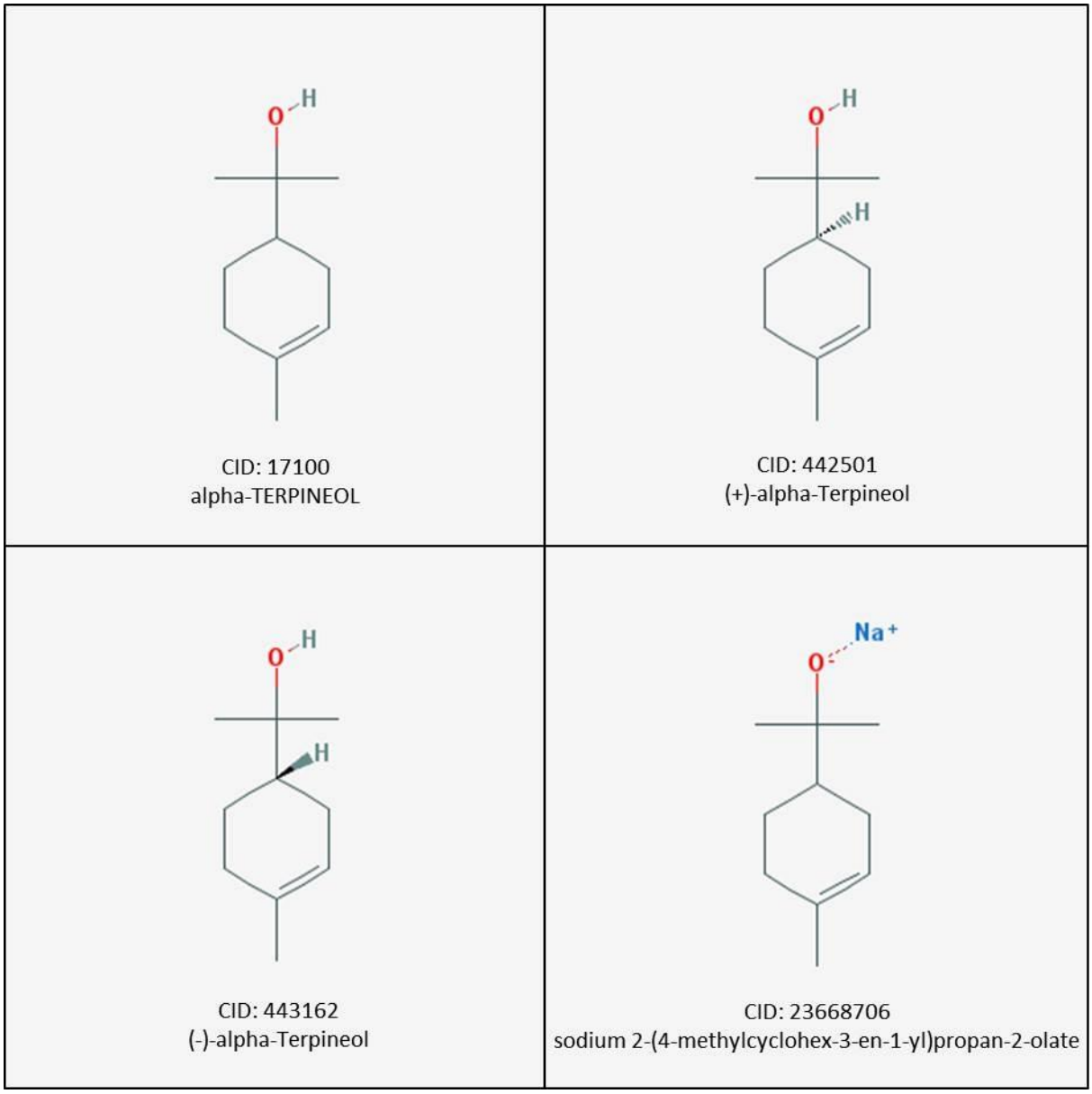


Figure 14 Example of homonymy in chemical naming

Chemical homonymy occurs when the same name can refer to different chemicals. Terpineol, its stereoisomers, and its sodium salt each have a different CID in PubChem but share common synonyms. Therefore, the same chemical name can match more than one PubChem CID. In cases of chemical homonymy, the stereoisomers and salts are mapped to the generic CID. These images were taken from PubChem (Kim et al., 2016).

Table 14 Homonymy among the DODSON chemicals (taken from Gabb and Blake, 2016a)

CID	Chemical Name	# Synonyms	# Homonymic CIDs	# Synonyms Taking Homonymic CIDs into Account
5355130	octinoxate	88	3	99
8572	benzophenone-1	107	1	109
5371084	methyl ionone	64	2	116
6448	isobornyl acetate	91	10	234
17100	terpineol	119	3	191
6549	linalool	118	2	197
22311	limonene	253	2	407
7585	alpha-hexylcinnamaldehyde	25	2	111
107	benzylacetate	170	1	215
5590	4-tert-octylphenol monoethoxylate	193	1	198
6623	bisphenol A	189	1	204
2347	benzyl butyl phthalate	117	1	119
8343	bis(2-ethylhexyl) phthalate	179	1	182
7184	butyl paraben	141	1	145

Fourteen of the 55 chemicals listed by (Dodson et al., 2012) had at least one homonymic CID. In some cases, this significantly increased the number of potential synonyms associated with the chemical name. For example, accounting for homonymy increases the number of alpha-hexylcinnamaldehyde synonyms from 25 to 111.

To account for homonymy, synonyms for a given chemical are compared to the synonyms of every other chemical in PubChem. If a match is found, the CIDs are considered to be homonymic. Fourteen of the 55 DODSON chemicals have at least one homonymic CID (Table 14). For example, the synthetic fragrance, methyl ionone (CID: 5371084), shares synonyms with two other chemicals: alpha-cetone (CID: 5372174) and 127-42-4 (CID: 16751505). The latter is a CAS-RN that is listed among the synonyms of both CIDs. In order to maximize coverage, the synonyms associated with all three CIDs are used when looking for methyl ionone among the consumer product ingredient lists.

Among all 11,964 target chemicals, 3,998 (33%) have at least one homonymic CID (the mean and median are 2.9 and 2 homonymic CIDs, respectively) (Supplemental Material, Homonymy Analysis): 1,538 have one homonymic CID; 1,023 have two; 493 have three; 294 have four; 185 have five; and 465 have six or more, including an extreme case of one target chemical having 88 homonymic CIDs. The latter is cyanocobalamin (CID: 184933), more commonly known as vitamin b12. This CID has only four synonyms in PubChem, but this number expands to 1,307 when homonymy is taken into account. However, the expansion in

the number of synonyms is less dramatic for most homonymic CIDs. The mean and median expansion are only 87 and 57 additional synonyms, respectively.

Chapter 5: Chemical Exposure and Retention Factors

A probabilistic approach is used to rank the target chemicals and chemical combinations. The algorithm is shown in Figure 15. Each cycle begins by selecting a model consumer from the Kantar dataset. A model consumer consists of a set of product categories representing that consumer's average daily usage pattern and the consumer's weight in the Kantar sample (Figure 8). The probabilities of occurrence are computed for each chemical based on each model consumer's product categories and the proportion of products in these categories that contain the target chemicals. Consider, for example, a consumer who uses soap, deodorant, mouthwash, and toothpaste in an average day, and a heatmap that shows chemical C is present in 30% of soap (S) products in the CPDB, 10% of deodorants (D), 15% of mouthwashes (M), and 0% of toothpastes (T). This is sufficient to compute the probabilities that C occurs in four (C_4), three (C_3), two (C_2), one (C_1), or none (C_0) of the consumer's daily-use products, as the following computation illustrates. In this example, there is a 53.5% chance that none of the products in this consumer usage pattern contain C:

$$Pr(C_0) = Pr(S-) * Pr(D-) * Pr(M-) * Pr(T-) = 0.7 * 0.9 * 0.85 * 1.0 = 0.535$$

There is a 38.3% chance that only one of the products in this consumer usage pattern contains C:

$$Pr(C_1^1) = Pr(S-) * Pr(D+) * Pr(M-) * Pr(T-) = 0.7 * 0.1 * 0.85 * 1.0 = 0.059$$
$$Pr(C_1^2) = Pr(S-) * Pr(D-) * Pr(M+) * Pr(T-) = 0.7 * 0.9 * 0.15 * 1.0 = 0.095$$
$$Pr(C_1^3) = Pr(S-) * Pr(D-) * Pr(M-) * Pr(T+) = 0.7 * 0.9 * 0.85 * 0.0 = 0.0$$
$$Pr(C_1^4) = Pr(S+) * Pr(D-) * Pr(M-) * Pr(T-) = 0.3 * 0.9 * 0.85 * 1.0 = 0.229$$
$$Pr(C_1) = Pr(C_1^1) + Pr(C_1^2) + Pr(C_1^3) + Pr(C_1^4) = 0.059 + 0.095 + 0.0 + 0.229 = 0.383$$

There is a 7.7% chance that any two of the products in this consumer usage pattern contains C:

$$Pr(C_2^1) = Pr(S-) * Pr(D+) * Pr(M+) * Pr(T-) = 0.7 * 0.1 * 0.15 * 1.0 = 0.011$$
$$Pr(C_2^2) = Pr(S+) * Pr(D+) * Pr(M-) * Pr(T-) = 0.3 * 0.1 * 0.85 * 1.0 = 0.025$$
$$Pr(C_2^3) = Pr(S+) * Pr(D-) * Pr(M+) * Pr(T-) = 0.3 * 0.9 * 0.15 * 1.0 = 0.041$$
$$Pr(C_2) = Pr(C_2^1) + Pr(C_2^2) + Pr(C_2^3) = 0.011 + 0.025 + 0.041 = 0.077$$

There is a 0.5% chance that three of the products in this consumer usage pattern contain C:

$$Pr(C_3) = Pr(S+) * Pr(D+) * Pr(M+) * Pr(T-) = 0.3 * 0.1 * 0.15 * 1.0 = 0.005$$

There is zero chance that all four of the products in this consumer usage pattern will contain C because C is not present in any toothpaste products.

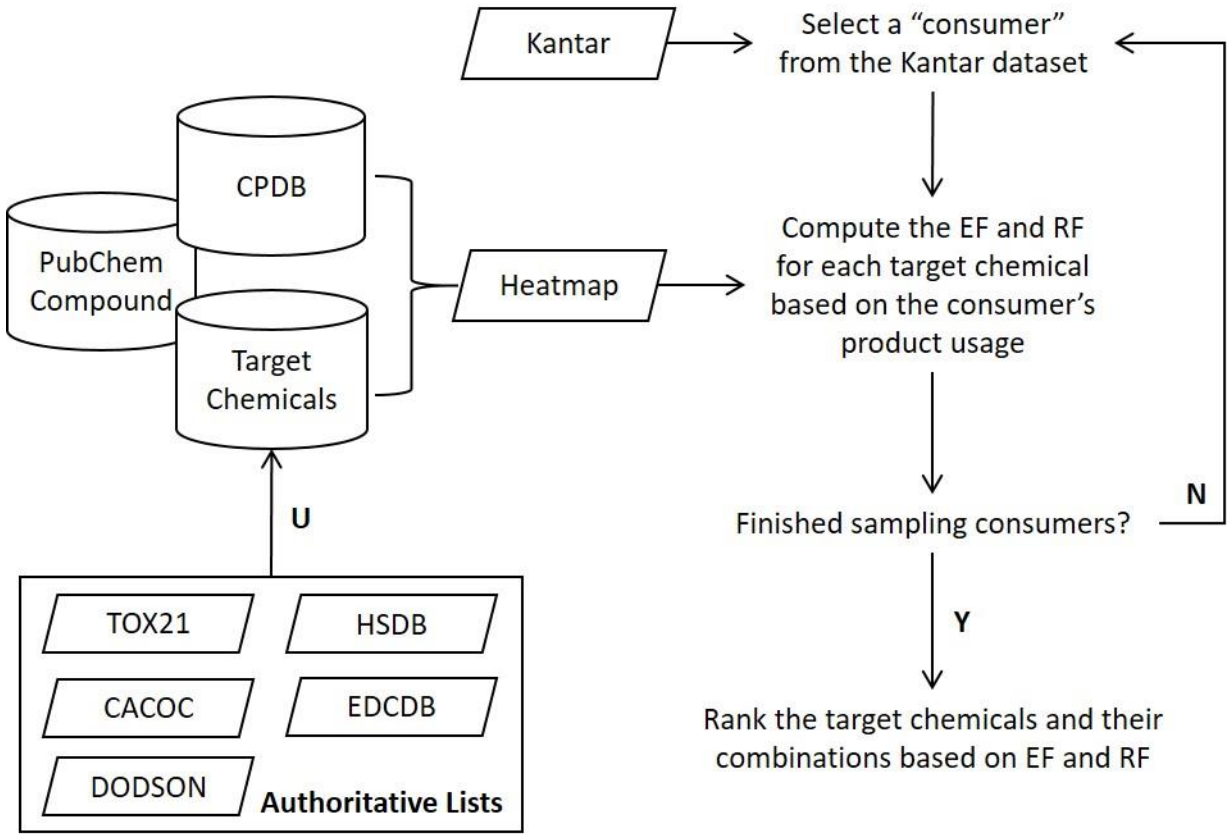


Figure 15 Probabilistic algorithm to prioritize the target chemicals and their combinations

This algorithm is used to find the most prevalent target chemicals and chemical combinations based on exposure factors (EF), retention factors (RF), and the product usage patterns of real consumers. The target chemicals consist of the union of the five authoritative lists: TOX21, HSDB, CACOC, EDCDB, and DODSON. CPDB refers to the database of consumer products and their ingredients and categories. The PubChem Compound database is used to resolve the target chemicals and product ingredients to unique identifiers. The heatmap contains the probability that a given product category contains a given target chemical.

The weighted exposure factor (EF) of chemical C for this consumer is calculated as:

$$EF_C = W \sum_{i=0}^{N} i * Pr(C_i)$$

where *N* is the number of product categories in the consumer's average daily usage pattern and *W* is the population weight of this consumer. If the consumer in the previous example represents 5.6% of the population, the weighted EF for C is:

$$EF_C = 0.056 * [(0.0)(0.535) + (1.0)(0.3835) + (2.0)(0.0765) + (3.0)(0.0045) + (4.0)(0.0)] = 0.056 * 0.55 = 0.031$$

The total exposure factor of each chemical over all consumers in the Kantar dataset is computed as follows:

$$Total\ EF_C = \sum_{j=1}^{M} \sum_{i=0}^{N_j} W_j * i * Pr(C_i)$$

where *M* is the number of model consumers and N_j is the number of products in a given consumer's daily usage pattern. Given the probability distribution of each target chemical by product category, and the population weight of each consumer, it is possible to rank the target chemicals and chemical combinations that occur most frequently in this consumer population. This rank serves as a proxy for likely exposure.

A separate set of factors is computed for each target chemical, taking the likely retention of chemicals into account based on the usage mode of consumer products. The exposure probability for a given chemical from a given product category is scaled by a retention factor (RF) taken from SCCS (2015). Products that are left on after application (makeup, moisturizer, antiperspirant/deodorant, and fragrance) have a RF of 1.0, and those that are rinsed off (cleanser, shampoo/conditioner, and hair removal products) have a RF of 0.01. Hair styling products have a RF of 0.1. Even though they are typically rinsed off after application, toothpaste and mouthwash have a slightly higher RF, 0.05 and 0.1 respectively, because of oral exposure and the possibility of ingestion. Therefore, when taking RF into account, the previous calculation of $Pr(C_3)$ becomes:

$$Pr(C_3) = RF_S * Pr(S+) * RF_D * Pr(D+) * RF_M * Pr(M+) * Pr(T-) = 0.01 * 0.3 * 1.0 * 0.1 * 0.1 * 0.15 * 1.0 = 4.5 \times 10^{-6}$$

The matrix of chemical proportion by product category, represented graphically as a heatmap (Figure 17), is a critical component of the EF and RF calculations and the subsequent prioritization schemes described in Chapter 6.2. The creation of this heatmap is described in Figure 13 and Table 11.

The approach described here is similar in spirit to that of Comiskey et al. (2015) but it has two important differences. First, all target chemicals are considered here instead of just the average concentration of generic fragrance by product category. Second, the algorithm described here estimates exposure based on the probability that a given chemical occurs in the product categories of model consumers. Exposure simulations are not used. In addition to these algorithmic differences, there are also differences in the product usage datasets. Comiskey et al. (2015) used a larger, more detailed subset of the Kantar Worldpanel. The present study uses a smaller and substantially cheaper subset of aggregate data. The larger, more expensive dataset contains individual consumers with demographic information rather than aggregate consumers, specific products rather than product categories, weeklong product usage diaries rather than averaged daily usage patterns, and 36,446 consumers in the U.S. and Europe instead of aggregate data from 11,000 American consumers. The EF- and RF-based computations described here are also similar to the frequent itemset mining approach that Kapraun et al. (2017) applied to NHANES, except that an exhaustive tabulation of every possible chemical combination to which a model consumer (i.e., a weighted set of product categories) could be exposed is considered, regardless of frequency. Product usage modes are also taken into account so that chemical combinations from different product sets can be weighted differently. For example, a chemical combination derived from products that are rinsed off after application would not carry as much weight as the same combination derived from products that are left on after application.

The exposure/retention computations described above assume that taking actual consumer product usage patterns into account affects how the target chemicals will be prioritized and adds value to the final prioritization. To test this assumption, exposure/retention is also computed for a “strawman” consumer—one who uses all product categories in the Kantar dataset with equal likelihood—to see if the prioritizations differ.

Chapter 6: Prevalent Target Chemicals in Consumer Products

This chapter contains the results of the informatics approach described in the previous chapters; namely, the various rankings for the target chemicals are presented here. The first set of results (Chapter 6.1) simply ranks the target chemicals and their combinations by prevalence within the consumer product sample and the personal care product subset. These results give some indication of which chemicals are most common in product formulations, though not necessarily the chemicals to which consumers are most likely to be exposed. Chapter 6.2 shows the results when consumer usage patterns are incorporated into the ranking scheme. These results indicate the target chemicals and chemical combinations that consumers are most likely to encounter in their everyday product usage.

The target chemicals are comprised of the union of five authoritative lists, as described in Chapter 3.1. No list is favored over another in this work. They are treated equally. However, Chapter 6.3 provides separate analyses for the individual lists in case there is specific interest in one of these sets of chemicals.

6.1 Ranking Prevalent Chemicals and Chemical Combinations in Consumer Products

The top 25 target chemicals detected in the complete product sample are shown in Table 15. The complete ranked list of detected chemicals is included in the Supplemental Material (Target Chemicals Detected). Most of these chemicals appear in more than one authoritative list. With the exception of octamethyltrisiloxane (an anti-foaming agent and skin conditioner), those that appear in three or more lists are also suspected EDCs: 2-phenoxyethanol (a common preservative), limonene and linalool (fragrances and flavorings), and methyl 4-hydroxybenzoate and propyl 4-hydroxybenzoate (parabens used as preservatives and/or fragrances).

Table 15 Twenty-five most prevalent target chemicals in the complete product sample

CID	Chemical Name	Number of Products Containing This Chemical	Percentage of Products Containing This Chemical	Authoritative List(s) Containing This Chemical
753	glycerol	13484	25.08977913	TOX21, HSDB
26042	titanium dioxide	11532	21.45767821	HSDB, CACOC
24261	silica	10717	19.94120164	TOX21, CACOC
311	citric acid	10438	19.42206427	TOX21, HSDB
86472	vitamin e acetate	8471	15.76205273	TOX21
14833	raphisiderite	7829	14.56747856	HSDB
24705	octamethyltrisiloxane	7677	14.28465102	TOX21, HSDB, CACOC
31236	2-phenoxyethanol	7588	14.11904806	TOX21, HSDB, DODSON
5234	sodium chloride	7338	13.6538712	HSDB
7456	methyl 4-hydroxybenzoate	5420	10.08503433	TOX21, HSDB, CACOC, EDCDB, DODSON
1030	1,2-propanediol	5326	9.910127831	TOX21, HSDB
7175	propyl 4-hydroxybenzoate	4756	8.849524589	TOX21, HSDB, CACOC, EDCDB
14985	vitamin e	4745	8.829056807	TOX21, HSDB
4678	dl-panthenol	4689	8.724857191	TOX21, HSDB
5281	stearic acid	4667	8.683921627	TOX21, HSDB
6049	edta	4460	8.298755187	TOX21, HSDB
2116	alpha-tocopherol	4268	7.941499358	HSDB
22311	limonene	4097	7.623318386	TOX21, HSDB, DODSON
17559	acid blue 9	3928	7.308858828	HSDB
441411	l-epinephrine hydrochloride	3870	7.200937797	HSDB
6549	linalool	3853	7.16930577	TOX21, HSDB, DODSON
14749	carminic acid	3321	6.179409411	TOX21, HSDB
517055	sodium benzoate	3237	6.023109986	TOX21, HSDB
23676745	potassium sorbate	3175	5.907746125	TOX21, HSDB
2682	1-hexadecanol	3080	5.730978918	TOX21, HSDB

The complete table of detected target chemicals is provided in the Supplemental Material (Target Chemicals Detected). The complete table of overlapping chemicals is provided in the Supplemental Material (Target Chemical Overlap).

The remaining top 25 target chemicals that appear in only one or two of the authoritative lists have a wide range of functions in consumer products: colorants (titanium oxide, raphisiderite, acid blue 9, carminic acid), preservatives (potassium sorbate, sodium benzoate, sodium chloride), humectants/skin conditioners (glycerol; 1,2-propanediol; dl-panthenol), surfactants/emulsifiers (1-hexadecanol, stearic acid), chelating agents (citric acid, edta), and anti-caking agents (silica). Note that many of these ingredients have multiple uses.

For example, titanium oxide is also used as an ultraviolet filter, sodium chloride is also used as a flavoring agent, and citric acid is also used as a preservative and flavoring agent. Tocopherols (used as antioxidants, preservatives, and/or skin conditioners) are common ingredients in consumer products but only appear in the TOX21 and HSDB lists. Interestingly, the lists disagree on the specific compounds. HSDB includes only vitamin e and alpha-tocopherol, while TOX21 includes vitamin e; vitamin e acetate, succinate, and nicotinate; delta- and gamma-tocopherol; and tocophersolan.

The heatmap of DODSON chemicals for the complete product sample is shown in Figure 16. The larger product sample does not change the distribution compared to the original heatmap in Gabb and Blake (2016a). It is worth noting that the antimicrobials, triclosan and triclocarban, are still present in the CPDB but they are relatively rare (Table 1 and Figure 16), which is not surprising as these chemicals are being phased out of consumer products due to increasing regulatory scrutiny (EPA, 2010, 2015a; FDA, 2016) and consumer pressure (APUA, 2011; Coleman-Lochner et al., 2014; EWG, 2014).

The distribution of the top 25 target chemicals among the ten product categories in the Kantar dataset is shown in Figure 17. [The complete heatmap used to compute EF and RF is included in the Supplemental Material (Heatmap).] Note that the top 25 chemicals are the same for the complete product sample (Table 15) and the subset of personal care products in the Kantar categories (Figure 17). Most of these chemicals appear in all ten product categories. Glycerol is a very common ingredient in oral care (toothpaste and mouthwash) and skin care (moisturizers, hair removal, and cleansers) products. Titanium oxide and raphisiderite are very common in makeup. As expected, the fragrance chemicals limonene and linalool are very common in fragrance products.

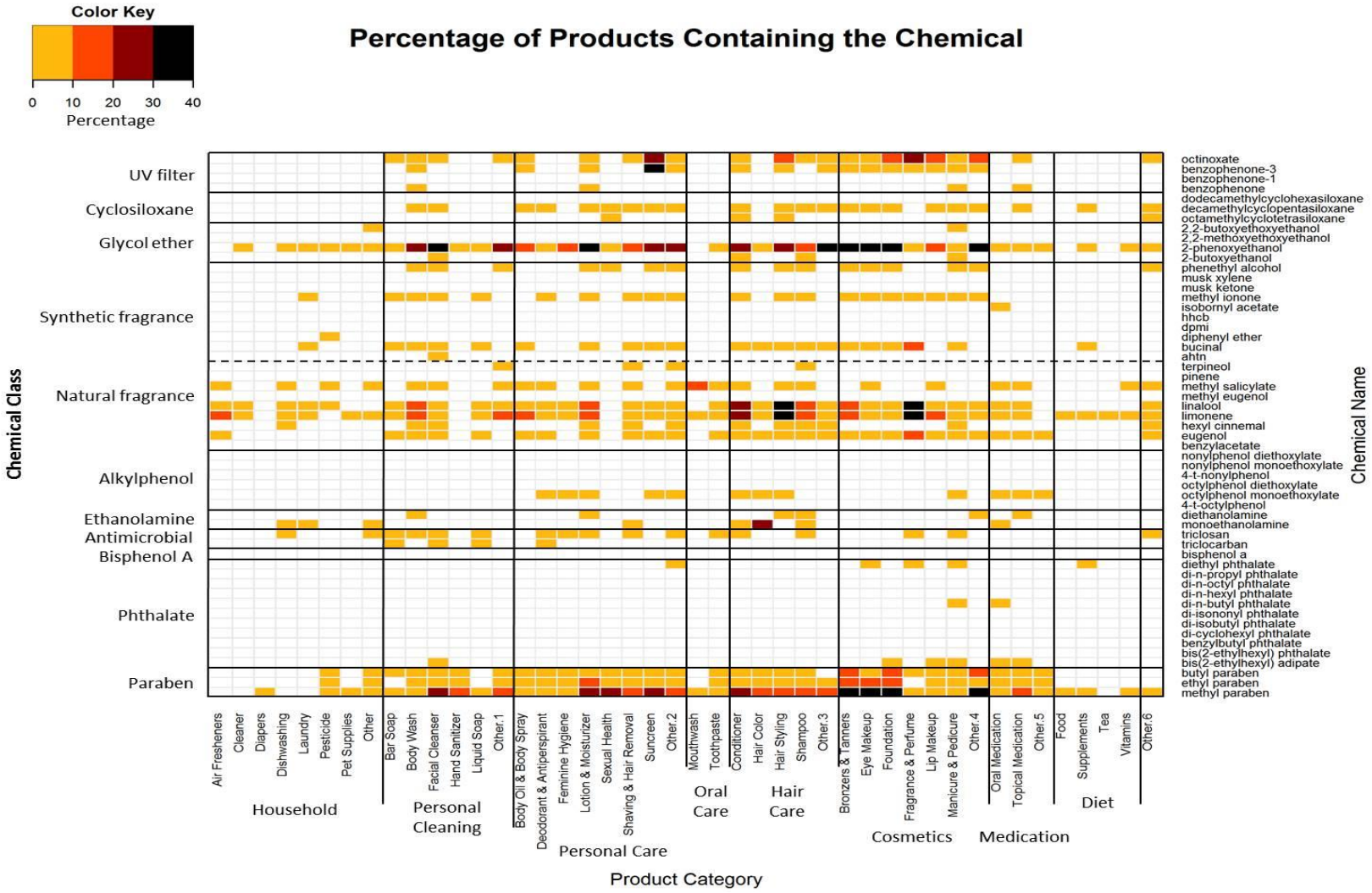


Figure 16 Heatmap of chemical prevalence by product category for the DODSON chemicals

This heatmap shows the prevalence of the DODSON chemicals in the complete product sample. Broad and specific consumer product categories are shown along the horizontal axis. Chemical class is shown on the left vertical axis and specific chemical ingredients are shown on the right vertical axis. White indicates that a chemical was not found in a product category. Yellow indicates that > 0 – 10% of the products in the category contain the chemical. Orange indicates that > 10 – 20% of the products contain the chemical. Dark red indicates that > 20 – 30% of the products contain the chemical. Black indicates that > 30 – 40% of the products contain the chemical.

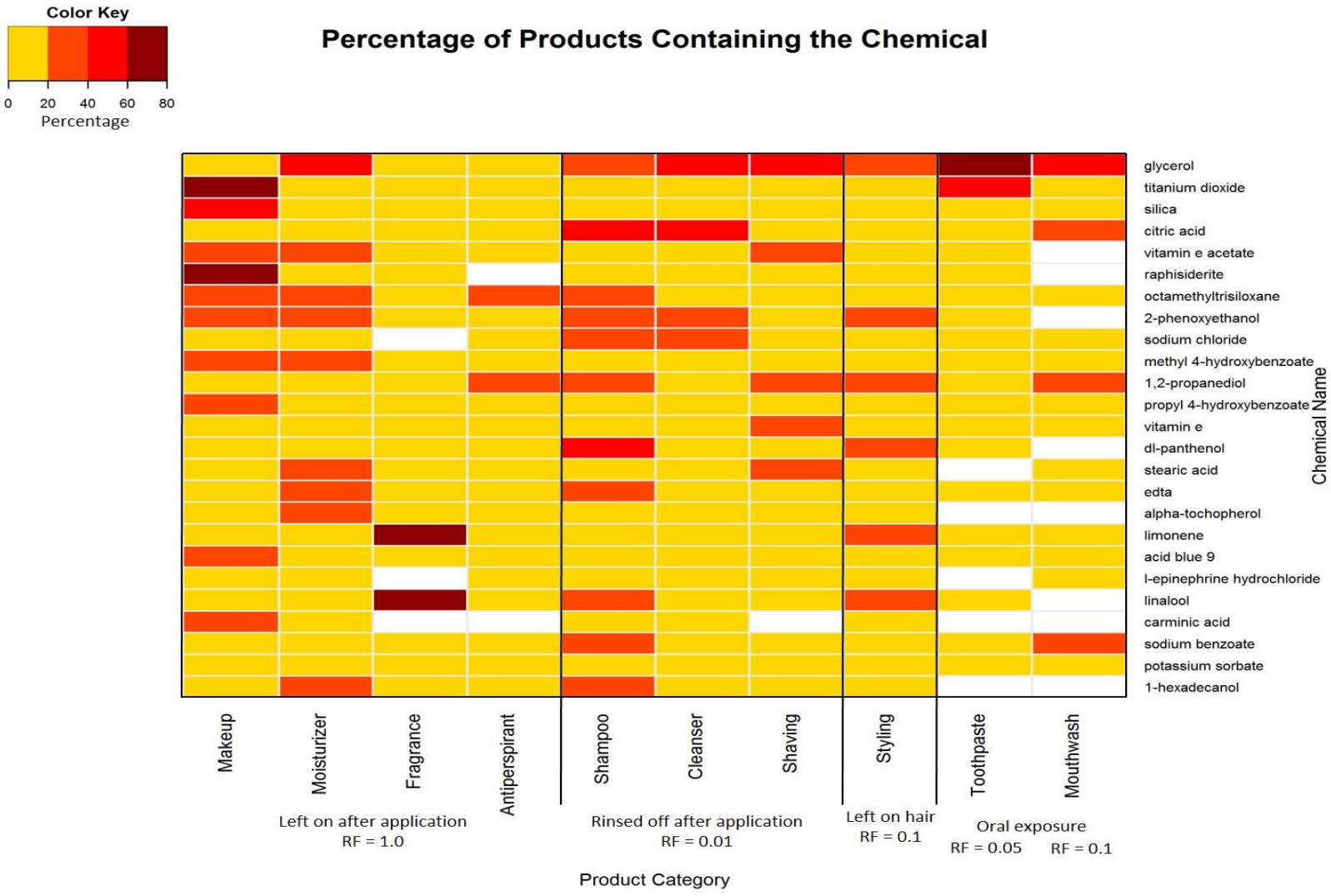


Figure 17 Heatmap showing prevalence by product category for the top-25 target chemicals

This heatmap shows the prevalence of the top-25 target chemicals in the product subset of the ten Kantar categories. White indicates that a chemical was not found in a product category. Yellow indicates that > 0 – 20% of the products in the category contain the chemical. Orange indicates that > 20 – 40% of the products contain the chemical. Red indicates that > 40 – 60% of the products contain the chemical. Dark red indicates that > 60 – 80% of the products contain the chemical.

It is important to note that the heatmap only shows the prevalence of chemicals by product category. It should not be used to draw conclusions about chemical safety. For example, one could speculate that alpha-tocopherol, carminic acid, 1-hexadecanol are absent from oral care products because they are harmful if ingested. However, it is equally plausible that these chemicals are absent because they provide no value to the product formulation. The heatmap is simply a tool to prioritize chemicals for risk assessment based on their prevalence in the consumer product sample, as was done in Gabb and Blake (2016a).

The combinatorial analysis of Gabb and Blake (2016a) tabulated chemical combinations co-occurring within individual products (i.e., per-product combinations) (Figure 18). A similar tabulation is done here but for a larger product sample (55,209 vs. 38,975) and a much larger set of target chemicals (11,964 vs. 55). Figure 19 shows the workflow to tabulate the per-product co-occurrence of the target chemicals. The components of this workflow are described in Table 16.

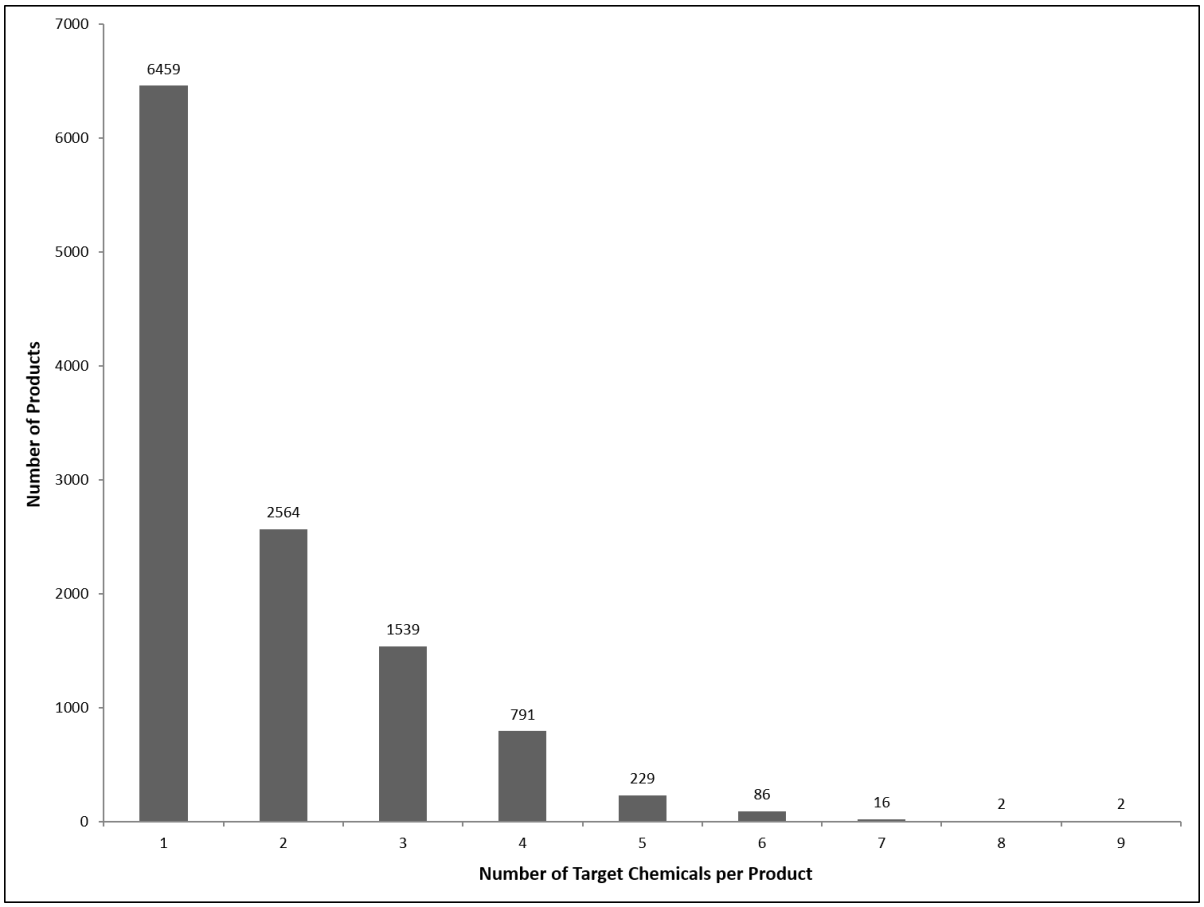


Figure 18 Number of products containing one or more DODSON chemicals

Of the 38,975 consumer products in the original product sample (Gabb and Blake, 2016a), 11,688 (30%) contain at least one of the potentially harmful chemicals identified by Dodson et al. (2012): 6459 contain only one target chemical, 2564 contain two, 1539 contain three, etc. Of the 11,688 products that contain a target chemical, 6,459 (55%) contain only one, while 5,229 (45%) contain more than one.

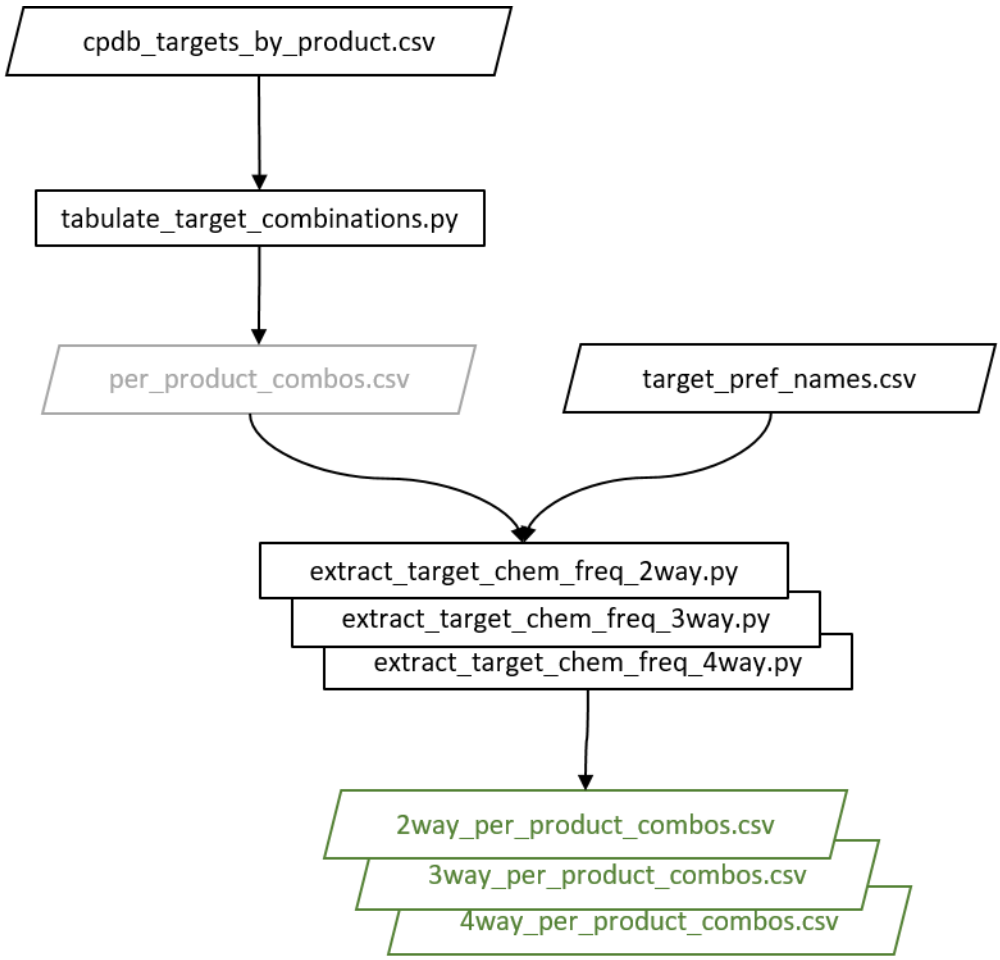


Figure 19 Tabulating per-product combinations of the target chemicals

Rectangles indicate computational processes, parallelograms indicate files, and arrows indicate data flow and dependencies. Green parallelograms indicate endpoints of the informatics workflow (i.e., the final results for a given set of target chemicals and consumer product sample). Gray parallelograms indicate intermediate or validation data that are not used in subsequent stages.

Table 16 Tabulating per-product combinations of the target chemicals

Computational Processes	
File Name	**Description**
tabulate_target_combinations.py	This program tabulates the combinations of target chemicals that occur in each consumer product in the database.
`Command: python tabulate_target_combinations.py < cpdb_targets_by_product.csv`	
extract_target_chem_freq_{2..4}way.py	These simple scripts extract the 2-, 3-, and 4-way per-product combinations of the target chemicals from the previous tabulation (i.e., `per_product_combos.csv`)

`Command: python extract_target_chem_freq_2way.py`
`python extract_target_chem_freq_3way.py`
`python extract_target_chem_freq_4way.py`	

Data	
File Name	**Description**
cpdb_targets_by_product.csv	See Table 11.
target_pref_names.csv	See Table 9.
per_product_combos.csv	Each record in this comma-delimited file describes a target chemical combination that occurs in the consumer products in the database. The first field is the number of chemicals in the combination (i.e., 2-, 3-, or 4-way), the second field contains the CIDs of the co-occurring chemicals, the third field is the number of products containing this combination, and the last field is the percentage of products containing this combination.
{2..4}way_per_product_combos.csv	These files have the same format as `per_product_combos.csv` but each contains only 2-, 3-, or 4-way combinations of the target chemicals.

Computational processes and files in the workflow to tabulate per-product combinations of the target chemicals. Backslashes indicate command-line continuation.

The 20 most common 2- and 3-way per-product combinations of the target chemicals are shown in Table 17 and Table 18. The complete ranked lists are included in the Supplemental Material: 2-way Combo (per-product) and 3-way Combo (per-product). It is not surprising that the most common chemicals in the consumer product sample (Table 15) tend to co-occur within the same product (Table 17 and Table 18), though several lower-ranked chemicals appear in the 3-way combinations (i.e., butyl acetate, ethyl acetate, nitrocellulose, isopropanol, and tin dioxide).

Table 17 Twenty most common 2-way per-product chemical combinations

Chemical 1	**Chemical 2**	**Number of Products Containing this Combination**
raphisiderite	titanium dioxide	7173
silica	titanium dioxide	5439
glycerol	2-phenoxyethanol	4149
raphisiderite	silica	4127
propyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate	3918
octamethyltrisiloxane	titanium dioxide	3689
citric acid	glycerol	3661
titanium dioxide	vitamin e acetate	3196
carminic acid	titanium dioxide	3184
glycerol	vitamin e acetate	3143
carminic acid	raphisiderite	3107
raphisiderite	octamethyltrisiloxane	3100
linalool	limonene	3019
vitamin e	vitamin e acetate	2994
acid blue 9	titanium dioxide	2865
glycerol	octamethyltrisiloxane	2829
silica	octamethyltrisiloxane	2779
raphisiderite	vitamin e acetate	2707
citric acid	sodium chloride	2690
octamethyltrisiloxane	2-phenoxyethanol	2660

Twenty most frequently occurring 2-way per-product combinations among the target chemicals. The complete table is provided in Supplemental Material: 2-way Combo (per-product).

Table 18 Twenty most common 3-way per-product chemical combinations

Chemical 1	Chemical 2	Chemical 3	Number of Products Containing this Combination
raphisiderite	silica	titanium dioxide	3993
carminic acid	raphisiderite	titanium dioxide	3026
raphisiderite	octamethyltrisiloxane	titanium dioxide	2910
raphisiderite	titanium dioxide	vitamin e acetate	2522
raphisiderite	acid blue 9	titanium dioxide	2506
silica	octamethyltrisiloxane	titanium dioxide	2339
isopropanol	ethyl acetate	butyl acetate	2174
isopropanol	butyl acetate	nitrocellulose	2140
raphisiderite	silica	octamethyltrisiloxane	2132
ethyl acetate	butyl acetate	nitrocellulose	2125
isopropanol	ethyl acetate	nitrocellulose	2071
carminic acid	silica	titanium dioxide	1967
raphisiderite	titanium dioxide	2-phenoxyethanol	1954
propyl 4-hydroxybenzoate	raphisiderite	titanium dioxide	1941
acid blue 9	silica	titanium dioxide	1913
carminic acid	raphisiderite	silica	1902
raphisiderite	titanium dioxide	tin dioxide	1892
raphisiderite	titanium dioxide	bismuth oxychloride	1848
titanium dioxide	butyl acetate	nitrocellulose	1834
silica	titanium dioxide	tin dioxide	1833

Twenty most frequently occurring 3-way per-product combinations among the target chemicals. The complete table is provided in Supplemental Material: 3-way Combo (per-product).

6.2 Ranking Prevalent Chemicals and Chemical Combinations among Consumers

6.2.1 Ranking Chemicals and Chemical Combinations by EF and RF

Knowing the target chemical combinations that occur within individual products is useful and interesting, but this only tells part of the overall combinatorial exposure story. Consumers who regularly use multiple consumer products are exposed to many more chemicals and chemical combinations. Figure 20 shows the workflow to tabulate per-consumer combinations of the target chemicals and compute exposure and retention factors based on product usage patterns of actual consumers, as described in Chapter 5. The components of this workflow are described in Table 19.

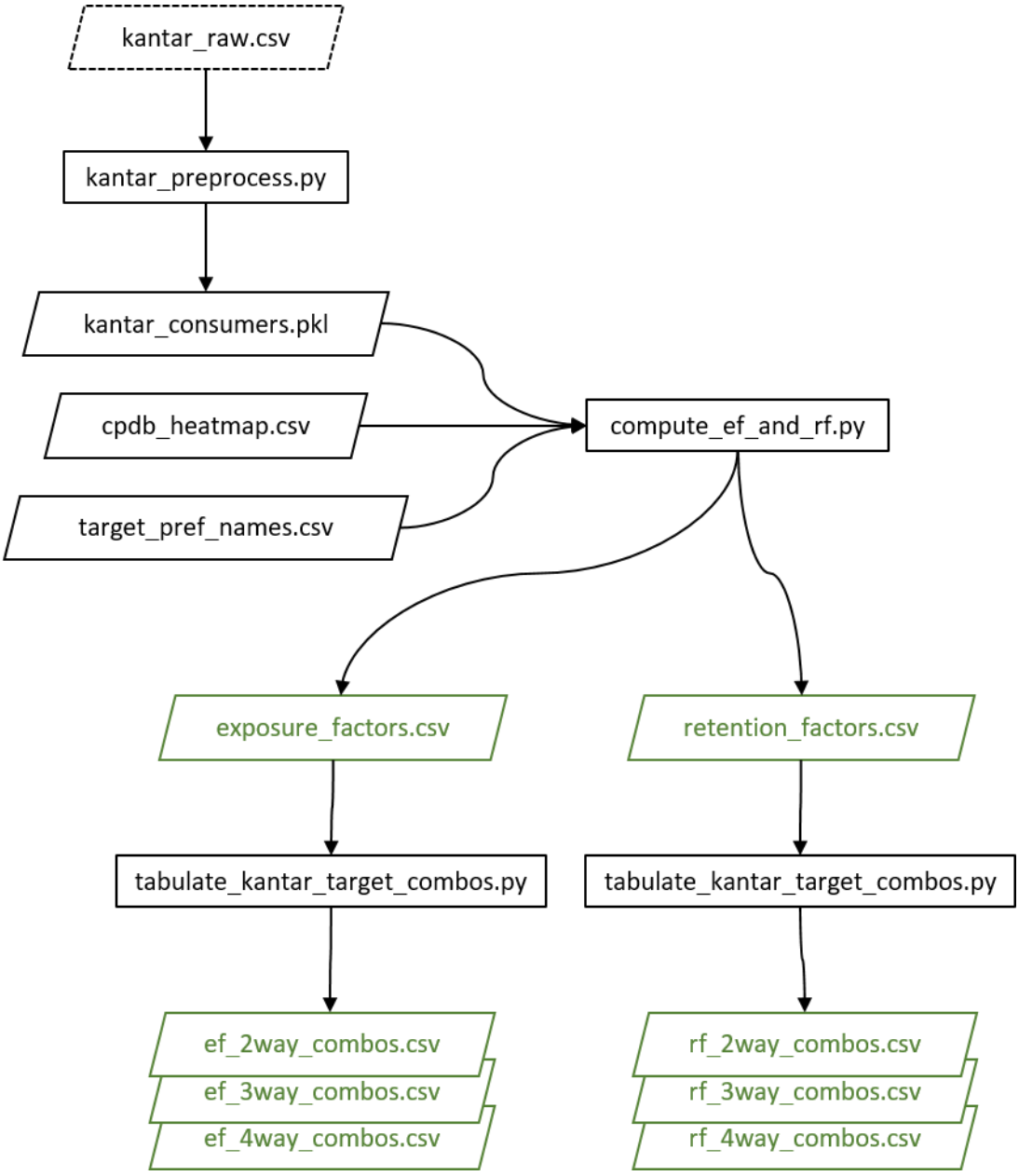


Figure 20 Computing per-consumer EF and RF of the target chemicals and their combinations

Rectangles indicate computational processes, parallelograms indicate files, and arrows indicate data flow and dependencies. The dotted line around the raw Kantar Worldpanel file indicates that these data were purchased from an external source. Green parallelograms indicate endpoints of the informatics workflow (i.e., the final results for a given set of target chemicals, model consumers, and consumer product sample).

Table 19 Tabulating per-consumer combinations of the target chemicals

Computational Processes	
File Name	**Description**
kantar_preprocess.py	This script extracts the necessary information from the raw Kantar Worldpanel data and assigns unique identifiers and weights to each model consumer and maps the Kantar product category codes to CPDB categories.
`Command: python kantar_preprocess.py`	
compute_ef_and_rf.py	This workhorse program implements the EF and RF equations described in Chapter 5 to compute the exposure and retention factors for each target chemical based on the previously computed heatmap and the model consumers from the Kantar Worldpanel.
`Command: python compute_ef_and_rf.py`	
tabulate_kantar_target_combos.py	This program computes the combinatorial probabilities described in Chapter 5 using the previously computed exposure and retention factors for each target chemical.
`Command: python tabulate_kantar_target_combos.py < exposure_factors.csv` `python tabulate_kantar_target_combos.py < retention_factors.csv`	
Data	
File Name	**Description**
kantar_raw.csv	This dataset was purchased from Kantar Worldpanel. It is described in Chapter 3.
kantar_consumers.pkl	Each record in this Python pickle file contains a model consumer from the Kantar Worldpanel dataset. The first field is an arbitrary consumer ID, the statistical weight for that consumer, and the list of product categories used by that consumer.
cpdb_heatmap.csv	See Table 11.
target_pref_names.csv	See Table 9.

Table 19 (cont.)

exposure_factors.csv, retention_factors.csv	Each record in these pipe-delimited files contains the computed exposure or retention factor for a target chemical. The first field is the CID, the second field is the preferred name, and the last field is the factor.
ef_{2..4}way_combos.csv, rf_{2..4}way_combos.csv	These pipe-delimited files contain the probabilities of each target chemical combination. Each record consists of the names of the chemicals in the combination followed by its probability based on the computed EF or RF.

Computational processes and files in the workflow to tabulate per-consumer combinations of the target chemicals. Backslashes indicate command-line continuation.

The present work improves upon Gabb and Blake (2016a) by ranking the combinations that occur among the mix of personal care products used by consumers (i.e., per-consumer combinations), informed by the Kantar dataset. However, unlike individual products, each model consumer is a weighted composite rather than an individual consumer, so chemical combinations are ranked using EF or RF. More specifically, for each model consumer, the set of all products in that consumer's usage pattern is generated. Next, the set of all target chemicals in these products is generated. Finally, the weighted probabilities (i.e., the product of EF_Cs for that combination multiplied by the weight of the model consumer) for all 2-, 3-, and 4-way chemical combinations within this set are computed. This was done across all model consumers to get the ranked lists of chemical combinations. To rank combinations based on likely retention, RF_Cs are used instead of EF_Cs. Table 20 and Table 21 show the 20 most common 2- and 3-way per-consumer combinations of the target chemicals, ranked by the EF product of each combination in the model consumer's product set.

Table 22 and Table 23 show the per-consumer combinations ranked by the RF product of each combination in the model consumer's product set. The complete lists of 2-, 3-, and 4-way chemical combinations are included in the Supplemental Material: 2-way Combo (EF, Kantar), 3-way Combo (EF, Kantar), 4-way Combo (EF, Kantar), 2-way Combo (RF, Kantar), 3-way Combo (RF, Kantar), and 4-way Combo (RF, Kantar). It is not surprising that the most common chemicals in the consumer product sample (Table 15) tend to co-occur within the

same product (Table 17 and Table 18). However, per-product rankings do not take usage frequency into account. Ranking chemical combinations based on the product combinations used by actual consumers (i.e., using the Kantar profiles) provides a more accurate method of assessing likely exposure because frequently used product categories are given more weight (Table 20 and Table 21). Similarly, taking product usage mode (i.e., left on after application, rinsed off after application, left on hair after application, and oral exposure) into account gives a more accurate assessment of likely retention (Cowan-Ellsberry and Robison, 2009; SCCS, 2015; Comiskey et al., 2015). Here, chemicals in products that are left on after application were given more weight when ranking combinations (Table 22 and Table 23).

Table 20 Top-20 2-way per-consumer chemical combinations ranked by EF

Chemical 1	**Chemical 2**
citric acid	glycerol
titanium dioxide	glycerol
glycerol	1,2-propanediol
hexitol	glycerol
saccharin	glycerol
dodecyl hydrogen sulfate	glycerol
octamethyltrisiloxane	glycerol
2-phenoxyethanol	glycerol
vitamin e acetate	glycerol
sodium chloride	glycerol
sodium benzoate	glycerol
silica	glycerol
octadecan-1-ol	glycerol
edta	glycerol
limonene	glycerol
dl-panthenol	glycerol
citric acid	titanium dioxide
glycerol	linalool
sodium hydroxide	glycerol
citric acid	1,2-propanediol

The complete table is provided in Supplemental Material: 2-way Combo (EF, Kantar).

Table 21 Top-20 3-way per-consumer chemical combinations ranked by EF

Chemical 1	**Chemical 2**	**Chemical 3**
citric acid	titanium dioxide	glycerol
citric acid	glycerol	1,2-propanediol
citric acid	hexitol	glycerol
titanium dioxide	glycerol	1,2-propanediol
citric acid	saccharin	glycerol
hexitol	titanium dioxide	glycerol
titanium dioxide	saccharin	glycerol
citric acid	dodecyl hydrogen sulfate	glycerol
hexitol	glycerol	1,2-propanediol
citric acid	octamethyltrisiloxane	glycerol
saccharin	glycerol	1,2-propanediol
citric acid	2-phenoxyethanol	glycerol
citric acid	vitamin e acetate	glycerol
citric acid	sodium chloride	glycerol
titanium dioxide	dodecyl hydrogen sulfate	glycerol
hexitol	saccharin	glycerol
octamethyltrisiloxane	titanium dioxide	glycerol
dodecyl hydrogen sulfate	glycerol	1,2-propanediol
2-phenoxyethanol	titanium dioxide	glycerol
titanium dioxide	vitamin e acetate	glycerol

The complete table is provided in the Supplemental Material: 3-way Combo (EF, Kantar).

Table 22 Top-20 2-way per-consumer chemical combinations ranked by RF

Chemical 1	**Chemical 2**
octamethyltrisiloxane	glycerol
octamethyltrisiloxane	octadecan-1-ol
octadecan-1-ol	glycerol
octamethyltrisiloxane	vitamin e acetate
octamethyltrisiloxane	1,2-propanediol
vitamin e acetate	glycerol
glycerol	1,2-propanediol
octamethyltrisiloxane	butylated hydroxytoluene
butylated hydroxytoluene	glycerol
octamethyltrisiloxane	silica
silica	glycerol
1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane
octamethyltrisiloxane	titanium dioxide
1-(2-butoxy-1-methylethoxy)propan-2-ol	glycerol
titanium dioxide	glycerol
2-phenoxyethanol	octamethyltrisiloxane
2-phenoxyethanol	glycerol
octadecan-1-ol	vitamin e acetate
octadecan-1-ol	1,2-propanediol
vitamin e acetate	1,2-propanediol

The complete table is provided in Supplemental Material: 2-way Combo (RF, Kantar).

Table 23 Top-20 3-way per-consumer chemical combinations ranked by RF

Chemical 1	Chemical 2	Chemical 3
octamethyltrisiloxane	octadecan-1-ol	glycerol
octamethyltrisiloxane	vitamin e acetate	glycerol
octamethyltrisiloxane	glycerol	1,2-propanediol
octamethyltrisiloxane	butylated hydroxytoluene	glycerol
octamethyltrisiloxane	silica	glycerol
1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	glycerol
octamethyltrisiloxane	titanium dioxide	glycerol
2-phenoxyethanol	octamethyltrisiloxane	glycerol
octamethyltrisiloxane	octadecan-1-ol	vitamin e acetate
octamethyltrisiloxane	octadecan-1-ol	1,2-propanediol
octadecan-1-ol	vitamin e acetate	glycerol
octadecan-1-ol	glycerol	1,2-propanediol
octamethyltrisiloxane	vitamin e acetate	1,2-propanediol
octamethyltrisiloxane	octadecan-1-ol	butylated hydroxytoluene
vitamin e acetate	glycerol	1,2-propanediol
octadecan-1-ol	butylated hydroxytoluene	glycerol
octamethyltrisiloxane	butylated hydroxytoluene	vitamin e acetate
octamethyltrisiloxane	octadecan-1-ol	silica
octamethyltrisiloxane	butylated hydroxytoluene	1,2-propanediol
butylated hydroxytoluene	vitamin e acetate	glycerol

The complete table is provided in Supplemental Material: 3-way Combo (RF, Kantar).

6.2.2 Qualitatively and Quantitatively Comparing the Ranked Lists

To determine the degree to which retention factors and/or consumer product usage patterns affect prioritization, it is necessary to compare the similarity of the ranked lists of the target chemicals described in Chapter 6.2.1. Table 24 provides a qualitative comparison of the EF and RF ranking schemes. Side-by-side comparison of the top 25 chemicals ranked by exposure factors taking consumer usage patterns into account (EF, Kantar), exposure factors ignoring usage patterns (EF, Strawman), retention factors taking usage patterns into account (RF, Kantar), and retention factors ignoring usage patterns (RF, Strawman); though only a snapshot of the highest-ranked chemicals, shows visually that the rankings change when likely retention and consumer usage patterns are taken into account. Many new chemicals enter the top 25 when actual product usage patterns and retention factors are taken into account: hexitol (surfactant), saccharin (flavoring), dodecyl hydrogen sulfate (surfactant), octadecan-1-ol (surfactant), sodium hydroxide (denaturant and pH balancer), sodium fluoride (anticaries

agent), butylated hydroxytoluene (antioxidant), 2-methyl-4-isothiazolin-3-one (preservative), edta tetrasodium (chelating agent), tween 20 (surfactant), 1-(2-butoxy-1-methylethoxy)propan-2-ol (solvent), ethylene (solvent), 1-docosanol (emulsifier), triethanolamine (surfactant and pH balancer), and geraniol (fragrance). The complete ranked lists are included in the Supplemental Material: EF_C (Kantar), EF_C (Strawman), RF_C (Kantar), and RF_C (Strawman).

Two main conclusions can be drawn from the side-by-side comparison. First, accounting for actual consumer usage affects chemical rankings. Second, taking relative retention into account affects chemical rankings. These conclusions may seem obvious, but it was necessary to confirm them empirically. If consumer usage patterns had not affected the ranking, we could dispense with the Kantar consumer usage patterns. Likewise, if retention had not affected the ranking, we could ignore product usage modes and dispense with the RF computations.

Table 24 Comparison of the top-25 chemicals ranked using four approaches

	EF, Kantar	EF, Strawman	RF, Kantar	RF, Strawman
1	glycerol	glycerol	octamethyltrisiloxane	glycerol
2	citric acid	citric acid	glycerol	octamethyltrisiloxane
3	titanium dioxide	1,2-propanediol	octadecan-1-ol	vitamin e acetate
4	1,2-propanediol	titanium dioxide	vitamin e acetate	linalool
5	hexitol	limonene	1,2-propanediol	titanium dioxide
6	saccharin	linalool	butylated hydroxytoluene	limonene
7	dodecyl hydrogen sulfate	vitamin e acetate	silica	2-phenoxyethanol
8	octamethyltrisiloxane	2-phenoxyethanol	1-(2-butoxy-1-methylethoxy)propan-2-ol	1,2-propanediol
9	2-phenoxyethanol	octamethyltrisiloxane	titanium dioxide	silica
10	vitamin e acetate	hexitol	2-phenoxyethanol	butylated hydroxytoluene
11	sodium chloride	dl-panthenol	linalool	raphisiderite
12	sodium benzoate	sodium benzoate	limonene	octadecan-1-ol
13	silica	saccharin	raphisiderite	stearic acid
14	octadecan-1-ol	methyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate
15	edta	silica	alpha-tocopherol	vitamin e
16	limonene	edta	edta	propyl 4-hydroxybenzoate
17	dl-panthenol	sodium chloride	vitamin e	alpha-tocopherol
18	linalool	octadecan-1-ol	stearic acid	citric acid
19	sodium hydroxide	propyl 4-hydroxybenzoate	propyl 4-hydroxybenzoate	edta
20	methyl 4-hydroxybenzoate	butylated hydroxytoluene	citric acid	triethanolamine
21	sodium fluoride	vitamin e	ethylene	acid blue 9
22	butylated hydroxytoluene	dodecyl hydrogen sulfate	1-hexadecanol	dl-panthenol
23	2-methyl-4-isothiazolin-3-one	tween 20	1-docosanol	1-(2-butoxy-1-methylethoxy)propan-2-ol
24	potassium sorbate	raphisiderite	acid blue 9	1-hexadecanol
25	edta tetrasodium	acid blue 9	dl-panthenol	geraniol

Chemicals with no background shading appear in all four ranked lists, though not necessarily in the same positions. Chemicals highlighted in green appear in only the EF lists. Chemicals highlighted in orange appear in only the RF lists. Chemicals highlighted in blue appear in three of the four lists. Chemicals highlighted in pink appear in only one list.

Rank-biased overlap (RBO) (Webber et al., 2010) is used to compute the similarity of these lists. This technique is designed to handle non-conjoint lists and lists of different lengths,

and to weight higher-ranked items more heavily. All of these points are important when comparing the ranked lists of chemicals. RBO is calculated as follows, where *L* is the longer list, *S* is the shorter list, *l* is the length of *L*, *s* is the length of *S*, *X* is the overlap of *L* and *S* at a particular depth in the lists, and *p* is the weight given to the higher ranked elements (small *p* weights higher ranks more heavily while a *p* of 1 ignores rank altogether):

$$RBO(L,S,l,s) = \frac{1-p}{p}\left(\sum_{d=1}^{l}\frac{X_d}{d}p^d + \sum_{d=s+1}^{l}\frac{X_s(d-s)}{s*d}p^d\right) + \left(\frac{X_l - X_s}{l} + \frac{X_s}{s}\right)p^l$$

An RBO of zero means the ranked lists are disjoint and one means the ranked lists are identical. RBO scores were computed using the software provided by Webber et al. (2010) with the default $p = 0.98$.

RBO provides a quantitative measure of similarity over all of the detected target chemicals (not just the top 25). The RBO score of (EF, Kantar) versus (EF, Strawman) is 0.85, and (EF, Kantar) versus (RF, Kantar) is 0.68. An RBO score of 1.0 indicates identical rankings, so the scores confirm the qualitative inspection of the top 25 rankings (Table 24); namely, accounting for consumer product usage patterns and product usage modes affects chemical ranking, presumably making prioritization more rational. A complete set of RBO scores under a variety of conditions is included in the Supplemental Material (Rank-Biased Overlap).

6.3 Rankings for Each Authoritative List

6.3.1 Ranking the Tox21 10K Library

The heatmap of TOX21 chemicals among personal care products shows that this list covers much of the ingredient space (Figure 21). There are hotspots in all product categories, and 14 of top 25 chemicals are detected in every category. Glycerol, a multipurpose chemical, is common (>20%) in most categories. The fragrance chemicals, linalool and limonene, are detected in nearly every product category but are particularly common in fragrance (i.e., perfumes), shampoo, and hair styling products. The sweetener, hexitol, is detected in every category but it is particularly common in oral care products (toothpaste and mouthwash). Vitamin e and vitamin e acetate are both common, particularly in products applied to the skin. The same is true for the parabens, methyl and propyl 4-hydroxybenzoate. The emollient,

octamethyltrisiloxane, and the preservative, 2-phenoxyethanol, are detected in nearly every product category, with hotspots in several.

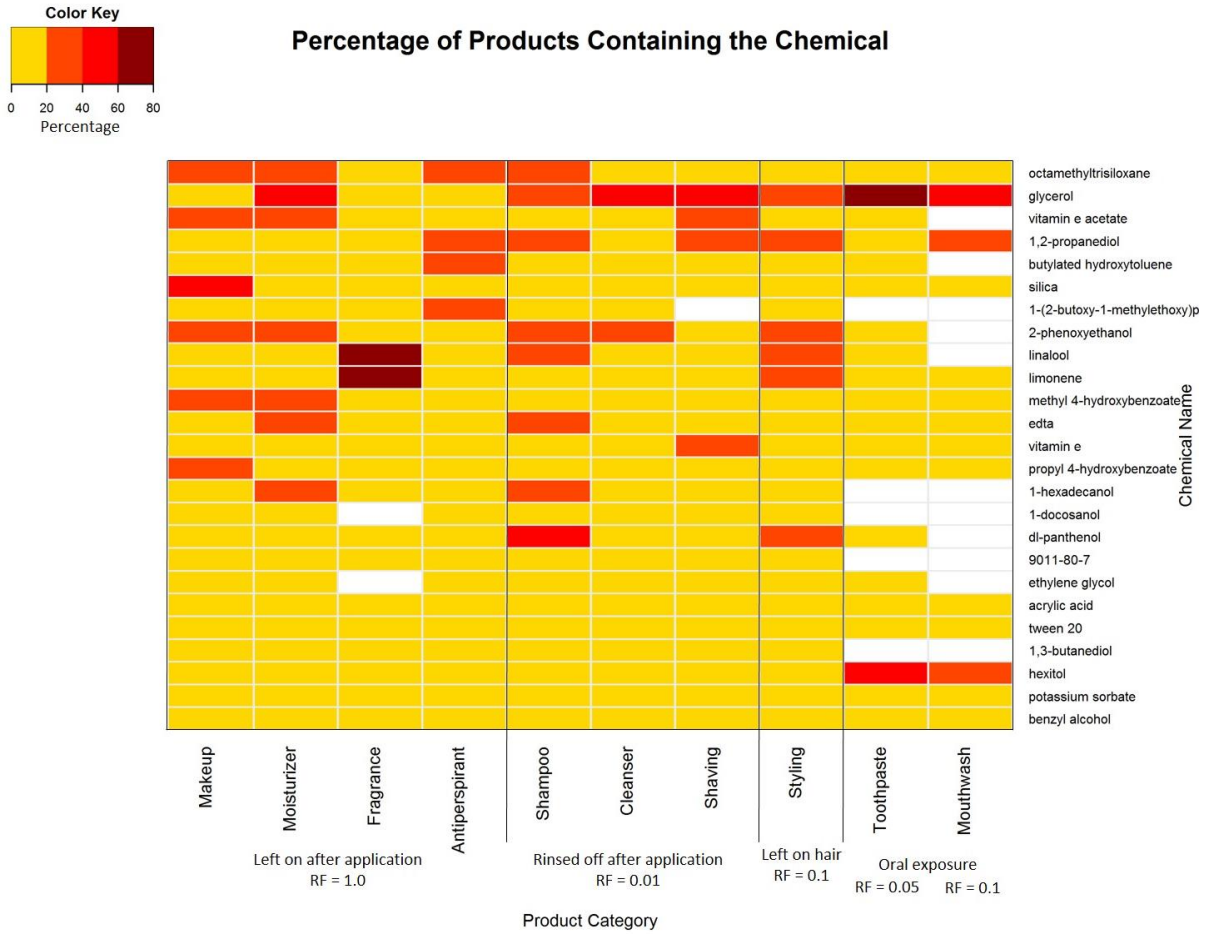


Figure 21 Heatmap of prevalence by product category for the top-25 TOX21 chemicals

The chemicals are ranked top to bottom by RF score. White indicates that a chemical was not found in a product category. Yellow indicates that > 0 – 20% of the products in the category contain the chemical. Orange indicates that > 20 – 40% of the products contain the chemical. Red indicates that > 40 – 60% of the products contain the chemical. Dark red indicates that > 60 – 80% of the products contain the chemical.

There are no surprises in the top 2- and 3-way combinations for TOX21 (Table 25 and Table 26). The same high-RF chemicals in Figure 21 comprise the highest ranked combinations.

Table 25 Top-25 2-way TOX21 chemical combinations ranked by RF

	Chemical 1	**Chemical 2**
1	octamethyltrisiloxane	glycerol
2	octamethyltrisiloxane	vitamin e acetate
3	octamethyltrisiloxane	1,2-propanediol
4	vitamin e acetate	glycerol
5	1,2-propanediol	glycerol
6	octamethyltrisiloxane	butylated hydroxytoluene
7	glycerol	butylated hydroxytoluene
8	octamethyltrisiloxane	silica
9	silica	glycerol
10	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane
11	1-(2-butoxy-1-methylethoxy)propan-2-ol	glycerol
12	2-phenoxyethanol	octamethyltrisiloxane
13	2-phenoxyethanol	glycerol
14	1,2-propanediol	vitamin e acetate
15	vitamin e acetate	butylated hydroxytoluene
16	1,2-propanediol	butylated hydroxytoluene
17	silica	vitamin e acetate
18	1,2-propanediol	silica
19	silica	butylated hydroxytoluene
20	octamethyltrisiloxane	linalool
21	glycerol	linalool
22	octamethyltrisiloxane	limonene
23	1-(2-butoxy-1-methylethoxy)propan-2-ol	vitamin e acetate
24	limonene	glycerol
25	1-(2-butoxy-1-methylethoxy)propan-2-ol	1,2-propanediol

Table 26 Top-25 3-way TOX21 chemical combinations ranked by RF

	Chemical 1	**Chemical 2**	**Chemical 3**
1	octamethyltrisiloxane	vitamin e acetate	glycerol
2	octamethyltrisiloxane	1,2-propanediol	glycerol
3	octamethyltrisiloxane	glycerol	butylated hydroxytoluene
4	octamethyltrisiloxane	silica	glycerol
5	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	glycerol
6	2-phenoxyethanol	octamethyltrisiloxane	glycerol
7	octamethyltrisiloxane	1,2-propanediol	vitamin e acetate
8	1,2-propanediol	vitamin e acetate	glycerol
9	octamethyltrisiloxane	vitamin e acetate	butylated hydroxytoluene
10	octamethyltrisiloxane	1,2-propanediol	butylated hydroxytoluene
11	vitamin e acetate	glycerol	butylated hydroxytoluene
12	1,2-propanediol	glycerol	butylated hydroxytoluene
13	octamethyltrisiloxane	silica	vitamin e acetate
14	octamethyltrisiloxane	1,2-propanediol	silica
15	silica	vitamin e acetate	glycerol
16	1,2-propanediol	silica	glycerol
17	octamethyltrisiloxane	silica	butylated hydroxytoluene
18	silica	glycerol	butylated hydroxytoluene
19	octamethyltrisiloxane	glycerol	linalool
20	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	vitamin e acetate
21	octamethyltrisiloxane	limonene	glycerol
22	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	1,2-propanediol
23	1-(2-butoxy-1-methylethoxy)propan-2-ol	vitamin e acetate	glycerol
24	1-(2-butoxy-1-methylethoxy)propan-2-ol	1,2-propanediol	glycerol
25	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	butylated hydroxytoluene

Even though TOX21 covers much of the personal care product ingredient space, this was not a specific consideration when the list was compiled (Tice et al., 2013, p. 757):

> *"The Tox21 Phase II compound library includes structurally defined compounds intended to broadly capture chemical and toxicological 'space.' The libraries include compounds with extensive to* no toxicological information *and with use, production, chemical class identity, and/or* environmental exposure patterns *that make them of potential concerns to regulatory agencies.* ... The physical cutoffs for the Phase II library were a molecular weight range of 100-1,000, a vapor pressure of < 10 Pa, and a ... desired solubility in [dimethyl sulfoxide] [of] 20 mM...*" (emphasis added)*

Existing evidence of toxicity was not a requirement for inclusion in TOX21, but environmental exposure was a criterion. Physical characteristics were also considered to ensure that the

selected chemicals were compatible with the HTS. Consequently, product ingredients that are low or high molecular weight, hydrophilic, and/or insoluble in dimethyl sulfoxide are not represented in TOX21. Silica (low molecular weight) is a notable exception, probably because of its high likelihood for environmental exposure.

The complete rankings are included in the Supplemental Material Individual Authoritative Lists. The individual chemicals ranked by EF and RF are in the TOX21 EF_C and TOX21 RF_C tables, respectively. The 2-, 3-, and 4-way combinations ranked by EF are in the TOX21 2-way Combo (EF, Kantar), TOX21 3-way Combo (EF, Kantar), and TOX21 4-way Combo (EF, Kantar) tables, respectively. The 2-, 3-, and 4-way combinations ranked by RF are in the TOX21 2-way Combo (RF, Kantar), TOX21 3-way Combo (RF, Kantar), and TOX21 4-way Combo (RF, Kantar) tables, respectively.

6.3.2 Ranking the Hazardous Substances Data Bank

It is not surprising that the HSDB heatmap (Figure 22) is similar to that of TOX21 (Figure 21) given the overlap between these lists (Figure 12). The only significant differences are acid blue 9, raphisiderite, and titanium dioxide, which occur in all personal care product categories, especially makeup. Acid blue 9 is hydrophilic and perhaps not sufficiently soluble in dimethyl sulfoxide to be included in TOX21. It is harder to speculate why raphisiderite and titanium dioxide were not included in TOX21. Their molecular weights are low but comparable to silica, which is included in TOX21.

Though TOX21 and HSDB overlap, their selection guidelines differ. As noted above, TOX21 was compiled with an eye toward HTS (Tice et al., 2013). This is not the case for HSDB (Fonger et al., 2014, pp. 210-211):

> *"Chemicals, drugs, dietary supplements, venoms, heavy metals and other candidate compounds are evaluated and selected by the HSDB chemical selection team, an internal NLM group. Candidate chemicals are nominated by members of NLM's staff, the public, scientific and regulatory agencies, and advisory groups. ... The selection team utilizes a rationale for chemical selection which includes: level of toxicity; human, animal, plant and aquatic exposure; amount of production and use and related factors such as regulatory status in the United States and other countries."*

The CPSC and FDA are among the regulatory agencies that nominate chemicals and provide toxicological information, so it is not surprising that the HSDB provides good coverage of the ingredient space of personal care products. Based on the selection criteria, however, it is still hard to see why water is included in the HSDB.

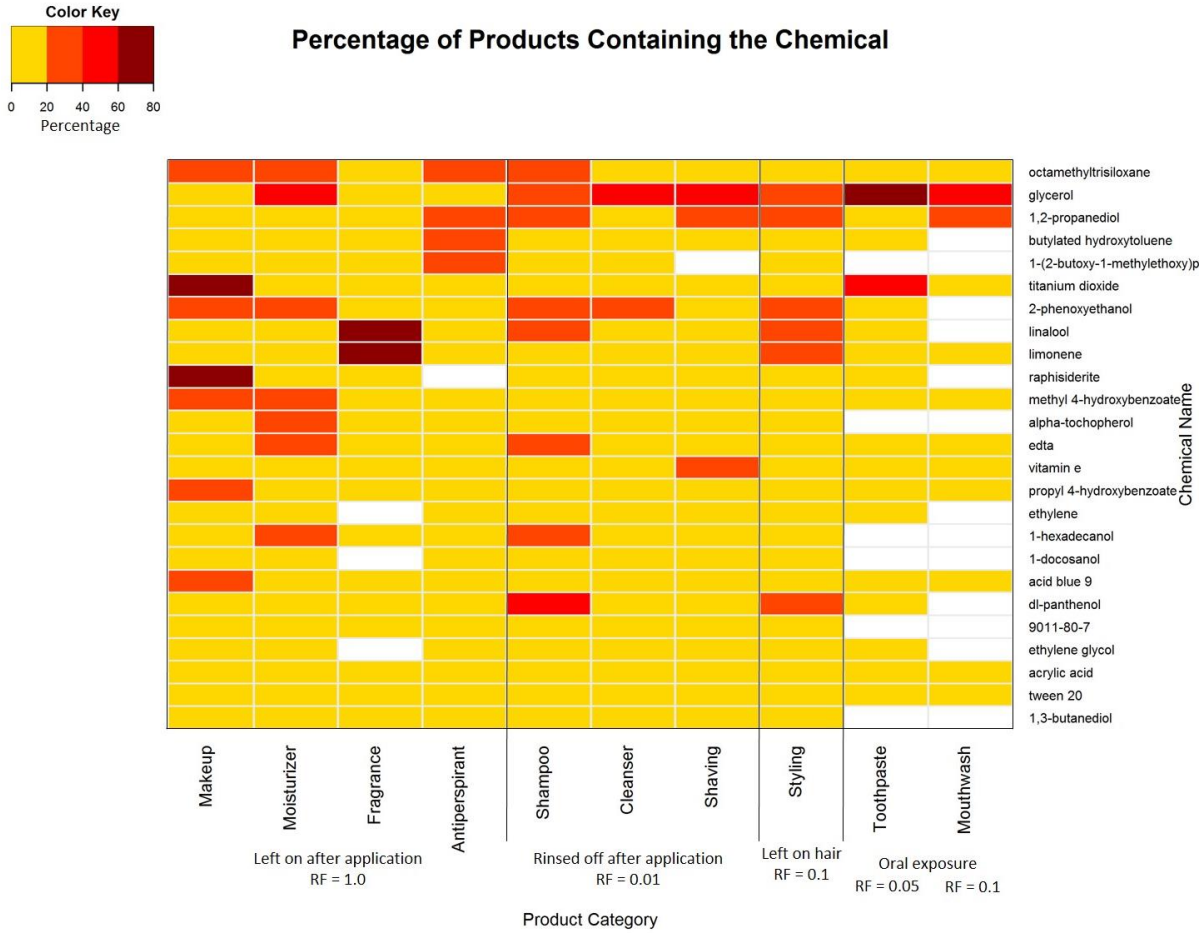


Figure 22 Heatmap of prevalence by product category for the top-25 HSDB chemicals

The chemicals are ranked top to bottom by RF score. White indicates that a chemical was not found in a product category. Yellow indicates that > 0 – 20% of the products in the category contain the chemical. Orange indicates that > 20 – 40% of the products contain the chemical. Red indicates that > 40 – 60% of the products contain the chemical. Dark red indicates that > 60 – 80% of the products contain the chemical.

As expected, the highest ranking HSDB chemicals in Figure 22 comprise the dominant 2- and 3-way combinations (Table 27 and Table 28).

Table 27 Top-25 2-way HSDB chemical combinations ranked by RF

	Chemical 1	**Chemical 2**
1	octamethyltrisiloxane	glycerol
2	octamethyltrisiloxane	1,2-propanediol
3	1,2-propanediol	glycerol
4	octamethyltrisiloxane	butylated hydroxytoluene
5	butylated hydroxytoluene	glycerol
6	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane
7	octamethyltrisiloxane	titanium dioxide
8	1-(2-butoxy-1-methylethoxy)propan-2-ol	glycerol
9	titanium dioxide	glycerol
10	2-phenoxyethanol	octamethyltrisiloxane
11	2-phenoxyethanol	glycerol
12	1,2-propanediol	butylated hydroxytoluene
13	octamethyltrisiloxane	linalool
14	glycerol	linalool
15	octamethyltrisiloxane	limonene
16	limonene	glycerol
17	1-(2-butoxy-1-methylethoxy)propan-2-ol	1,2-propanediol
18	1,2-propanediol	titanium dioxide
19	1-(2-butoxy-1-methylethoxy)propan-2-ol	butylated hydroxytoluene
20	titanium dioxide	butylated hydroxytoluene
21	2-phenoxyethanol	1,2-propanediol
22	octamethyltrisiloxane	raphisiderite
23	raphisiderite	glycerol
24	2-phenoxyethanol	butylated hydroxytoluene
25	1-(2-butoxy-1-methylethoxy)propan-2-ol	titanium dioxide

Table 28 Top-25 3-way HSDB chemical combinations ranked by RF

	Chemical 1	Chemical 2	Chemical 3
1	octamethyltrisiloxane	1,2-propanediol	glycerol
2	octamethyltrisiloxane	butylated hydroxytoluene	glycerol
3	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	glycerol
4	octamethyltrisiloxane	titanium dioxide	glycerol
5	2-phenoxyethanol	octamethyltrisiloxane	glycerol
6	octamethyltrisiloxane	1,2-propanediol	butylated hydroxytoluene
7	1,2-propanediol	butylated hydroxytoluene	glycerol
8	octamethyltrisiloxane	glycerol	linalool
9	octamethyltrisiloxane	limonene	glycerol
10	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	1,2-propanediol
11	octamethyltrisiloxane	1,2-propanediol	titanium dioxide
12	1-(2-butoxy-1-methylethoxy)propan-2-ol	1,2-propanediol	glycerol
13	1,2-propanediol	titanium dioxide	glycerol
14	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	butylated hydroxytoluene
15	octamethyltrisiloxane	titanium dioxide	butylated hydroxytoluene
16	2-phenoxyethanol	octamethyltrisiloxane	1,2-propanediol
17	1-(2-butoxy-1-methylethoxy)propan-2-ol	butylated hydroxytoluene	glycerol
18	titanium dioxide	butylated hydroxytoluene	glycerol
19	2-phenoxyethanol	1,2-propanediol	glycerol
20	octamethyltrisiloxane	raphisiderite	glycerol
21	2-phenoxyethanol	octamethyltrisiloxane	butylated hydroxytoluene
22	2-phenoxyethanol	butylated hydroxytoluene	glycerol
23	1-(2-butoxy-1-methylethoxy)propan-2-ol	octamethyltrisiloxane	titanium dioxide
24	1-(2-butoxy-1-methylethoxy)propan-2-ol	titanium dioxide	glycerol
25	1-(2-butoxy-1-methylethoxy)propan-2-ol	2-phenoxyethanol	octamethyltrisiloxane

The complete rankings are included in the Supplemental Material Individual Authoritative Lists. The individual chemicals ranked by EF and RF are in the HSDB EF_C and HSDB RF_C tables, respectively. The 2-, 3-, and 4-way combinations ranked by EF are in the HSDB 2-way Combo (EF, Kantar), HSDB 3-way Combo (EF, Kantar), and HSDB 4-way Combo (EF, Kantar) tables, respectively. The 2-, 3-, and 4-way combinations ranked by RF are in the HSDB 2-way Combo (RF, Kantar), HSDB 3-way Combo (RF, Kantar), and HSDB 4-way Combo (RF, Kantar) tables, respectively.

6.3.3 Ranking the California Chemicals of Concern

The goal of the Safer Consumer Products program of the California Department of Toxic Substances Control is to "reduce toxic chemicals in the products that consumers buy and use" (DTSC, 2016). As part of this effort, a list of chemicals "that exhibit a hazard trait and/or an environmental or toxicological endpoint" is maintained (DTSC, 2016). This list covers all categories of consumer products, not just personal care products. However, the CACOC does provide good coverage of the personal care product ingredient space, as the heatmap below demonstrates (Figure 23). Makeup ingredients, in particular, constitute the main hotspots among the CACOC. Four parabens are also among the top-25 chemicals: ethylparaben and methyl, propyl, and butyl 4-hydroxybenzoate. The emollient, octamethyltrisiloxane, is detected in nearly every product category, with hotspots in several. It is interesting to note that fragrance chemicals are not among the highest ranked chemicals. Fragrance, toothpaste, and mouthwash ingredients are only sparsely represented.

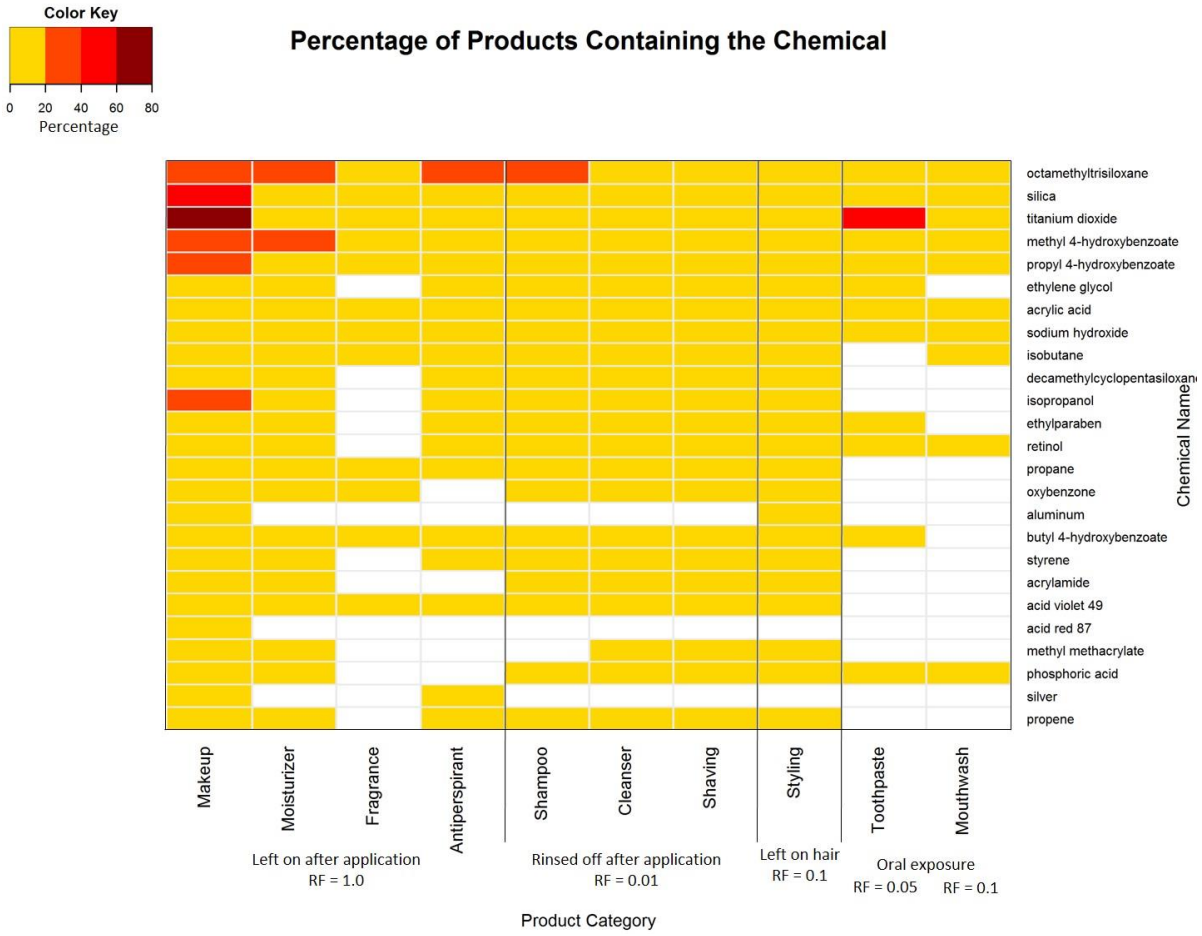


Figure 23 Heatmap of prevalence by product category for the top-25 CACOC chemicals

The chemicals are ranked top to bottom by RF score. White indicates that a chemical was not found in a product category. Yellow indicates that > 0 – 20% of the products in the category contain the chemical. Orange indicates that > 20 – 40% of the products contain the chemical. Red indicates that > 40 – 60% of the products contain the chemical. Dark red indicates that > 60 – 80% of the products contain the chemical.

As expected, the highest ranking CACOC chemicals in Figure 23 comprise the dominant 2- and 3-way combinations (Table 29 and Table 30).

Table 29 Top-25 2-way CACOC chemical combinations ranked by RF

	Chemical 1	**Chemical 2**
1	silica	octamethyltrisiloxane
2	octamethyltrisiloxane	titanium dioxide
3	silica	titanium dioxide
4	methyl 4-hydroxybenzoate	octamethyltrisiloxane
5	octamethyltrisiloxane	propyl 4-hydroxybenzoate
6	methyl 4-hydroxybenzoate	silica
7	methyl 4-hydroxybenzoate	titanium dioxide
8	silica	propyl 4-hydroxybenzoate
9	octamethyltrisiloxane	ethylene glycol
10	octamethyltrisiloxane	acrylic acid
11	titanium dioxide	propyl 4-hydroxybenzoate
12	octamethyltrisiloxane	sodium hydroxide
13	octamethyltrisiloxane	isobutane
14	silica	ethylene glycol
15	silica	acrylic acid
16	octamethyltrisiloxane	decamethylcyclopentasiloxane
17	octamethyltrisiloxane	isopropanol
18	ethylene glycol	titanium dioxide
19	acrylic acid	titanium dioxide
20	silica	sodium hydroxide
21	ethylparaben	octamethyltrisiloxane
22	methyl 4-hydroxybenzoate	propyl 4-hydroxybenzoate
23	titanium dioxide	sodium hydroxide
24	octamethyltrisiloxane	retinol
25	silica	isobutane

Table 30 Top-25 3-way CACOC chemical combinations ranked by RF

	Chemical 1	Chemical 2	Chemical 3
1	silica	octamethyltrisiloxane	titanium dioxide
2	methyl 4-hydroxybenzoate	silica	octamethyltrisiloxane
3	methyl 4-hydroxybenzoate	octamethyltrisiloxane	titanium dioxide
4	silica	octamethyltrisiloxane	propyl 4-hydroxybenzoate
5	octamethyltrisiloxane	titanium dioxide	propyl 4-hydroxybenzoate
6	methyl 4-hydroxybenzoate	silica	titanium dioxide
7	silica	octamethyltrisiloxane	ethylene glycol
8	silica	octamethyltrisiloxane	acrylic acid
9	silica	titanium dioxide	propyl 4-hydroxybenzoate
10	octamethyltrisiloxane	ethylene glycol	titanium dioxide
11	octamethyltrisiloxane	acrylic acid	titanium dioxide
12	silica	octamethyltrisiloxane	sodium hydroxide
13	methyl 4-hydroxybenzoate	octamethyltrisiloxane	propyl 4-hydroxybenzoate
14	octamethyltrisiloxane	titanium dioxide	sodium hydroxide
15	silica	octamethyltrisiloxane	isobutane
16	octamethyltrisiloxane	isobutane	titanium dioxide
17	silica	octamethyltrisiloxane	decamethylcyclopentasiloxane
18	silica	octamethyltrisiloxane	isopropanol
19	silica	ethylene glycol	titanium dioxide
20	silica	acrylic acid	titanium dioxide
21	octamethyltrisiloxane	decamethylcyclopentasiloxane	titanium dioxide
22	octamethyltrisiloxane	isopropanol	titanium dioxide
23	ethylparaben	silica	octamethyltrisiloxane
24	methyl 4-hydroxybenzoate	silica	propyl 4-hydroxybenzoate
25	silica	titanium dioxide	sodium hydroxide

The complete rankings are included in the Supplemental Material Individual Authoritative Lists. The individual chemicals ranked by EF and RF are in the CACOC EF_C and CACOC RF_C tables, respectively. The 2-, 3-, and 4-way combinations ranked by EF are in the CACOC 2-way Combo (EF, Kantar), CACOC 3-way Combo (EF, Kantar), and CACOC 4-way Combo (EF, Kantar) tables, respectively. The 2-, 3-, and 4-way combinations ranked by RF are in the CACOC 2-way Combo (RF, Kantar), CACOC 3-way Combo (RF, Kantar), and CACOC 4-way Combo (RF, Kantar) tables, respectively.

6.3.4 Ranking the Endocrine Disrupting Compounds Data Bank

The EDCDB chemicals (Montes-Grajales and Olivero-Verbel, 2015) are drawn from the European Union and Endocrine Disruption Exchange lists of potential endocrine disruptors (EU,

2017; TEDX, 2017) (http://eng.mst.dk/chemicals/chemicals-in-products/endocrine-disruptors/the-eu-list-of-potential-endocrine-disruptors/ and https://endocrinedisruption.org/interactive-tools/tedx-list-of-potential-endocrine-disruptors/search-the-tedx-list). The EDCDB catalogs chemicals that have documented *in vivo* or *in vitro* endocrine activity. The EDCDB heatmap below (Figure 24) is sparser than those of TOX21 (Figure 21), HSDB (Figure 22), and CACOC (Figure 23). This is not an indication of the relatively danger or safety of the chemicals in this list. Rather, it indicates that EDCDB does not provide the same level of coverage for the ingredient space of personal care products. Parabens (ethylparaben and methyl, propyl, and butyl 4-hydroxybenzoate) are among the top-ranked chemicals and are the only hotspots. Consequently, parabens dominate the highest ranking 2- and 3-way combinations of EDCDB chemicals (Table 31 and Table 32).

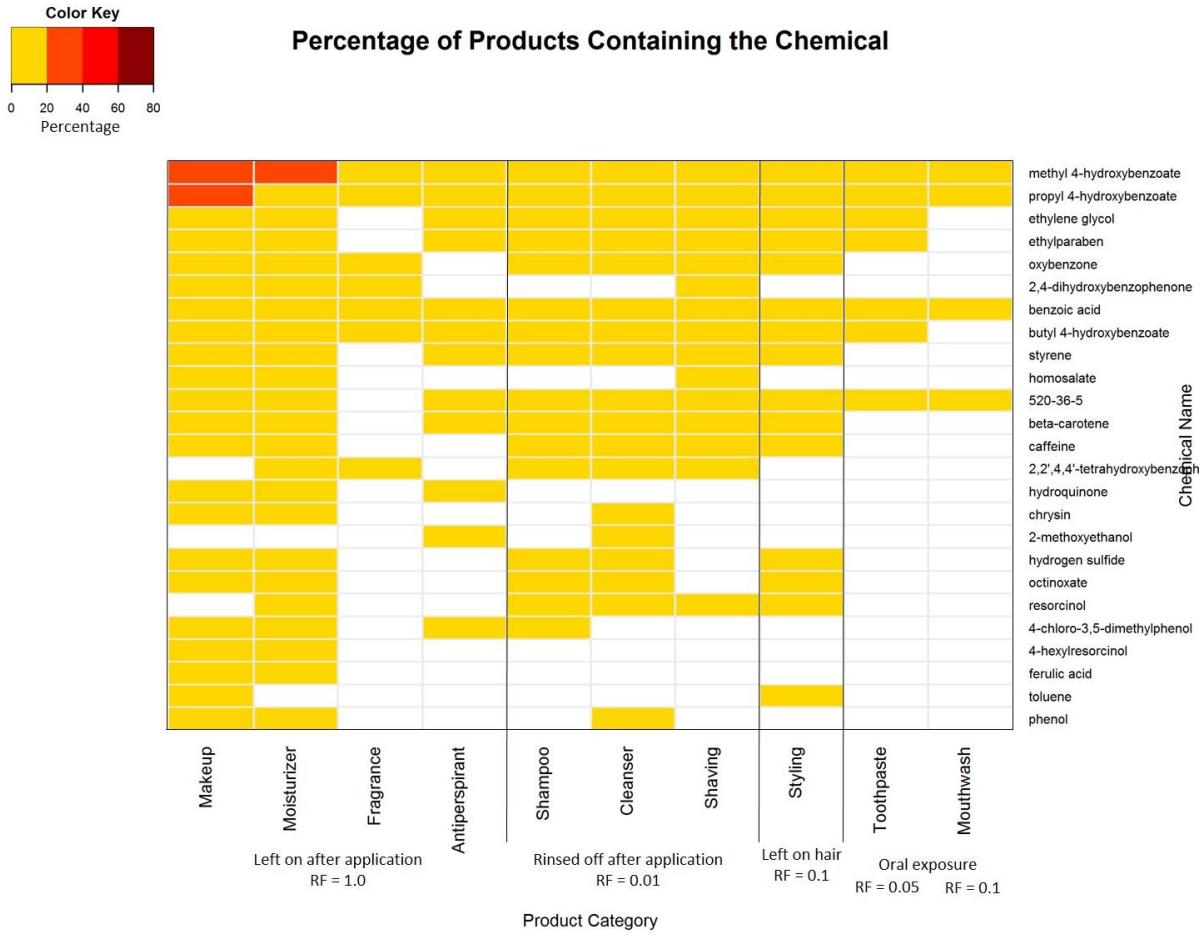


Figure 24 Heatmap of prevalence by product category for the top-25 EDCDB chemicals

The chemicals are ranked top to bottom by RF score. White indicates that a chemical was not found in a product category. Yellow indicates that > 0 – 20% of the products in the category contain the chemical. Orange indicates that > 20 – 40% of the products contain the chemical. Red indicates that > 40 – 60% of the products contain the chemical. Dark red indicates that > 60 – 80% of the products contain the chemical.

Table 31 Top-25 2-way EDCDB chemical combinations ranked by RF

	Chemical 1	**Chemical 2**
1	propyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate
2	ethylene glycol	methyl 4-hydroxybenzoate
3	propyl 4-hydroxybenzoate	ethylene glycol
4	methyl 4-hydroxybenzoate	ethylparaben
5	propyl 4-hydroxybenzoate	ethylparaben
6	oxybenzone	methyl 4-hydroxybenzoate
7	methyl 4-hydroxybenzoate	2,4-dihydroxybenzophenone
8	methyl 4-hydroxybenzoate	benzoic acid
9	methyl 4-hydroxybenzoate	butyl 4-hydroxybenzoate
10	propyl 4-hydroxybenzoate	oxybenzone
11	propyl 4-hydroxybenzoate	2,4-dihydroxybenzophenone
12	propyl 4-hydroxybenzoate	benzoic acid
13	ethylene glycol	ethylparaben
14	propyl 4-hydroxybenzoate	butyl 4-hydroxybenzoate
15	methyl 4-hydroxybenzoate	styrene
16	propyl 4-hydroxybenzoate	styrene
17	oxybenzone	ethylene glycol
18	homosalate	methyl 4-hydroxybenzoate
19	ethylene glycol	2,4-dihydroxybenzophenone
20	ethylene glycol	benzoic acid
21	ethylene glycol	butyl 4-hydroxybenzoate
22	520-36-5	methyl 4-hydroxybenzoate
23	homosalate	propyl 4-hydroxybenzoate
24	propyl 4-hydroxybenzoate	520-36-5
25	ethylene glycol	styrene

Table 32 Top-25 3-way EDCDB chemical combinations ranked by RF

	Chemical 1	Chemical 2	Chemical 3
1	propyl 4-hydroxybenzoate	ethylene glycol	methyl 4-hydroxybenzoate
2	propyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate	ethylparaben
3	propyl 4-hydroxybenzoate	oxybenzone	methyl 4-hydroxybenzoate
4	propyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate	2,4-dihydroxybenzophenone
5	propyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate	benzoic acid
6	ethylene glycol	methyl 4-hydroxybenzoate	ethylparaben
7	propyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate	butyl 4-hydroxybenzoate
8	propyl 4-hydroxybenzoate	ethylene glycol	ethylparaben
9	propyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate	styrene
10	oxybenzone	ethylene glycol	methyl 4-hydroxybenzoate
11	ethylene glycol	methyl 4-hydroxybenzoate	2,4-dihydroxybenzophenone
12	ethylene glycol	methyl 4-hydroxybenzoate	benzoic acid
13	ethylene glycol	methyl 4-hydroxybenzoate	butyl 4-hydroxybenzoate
14	propyl 4-hydroxybenzoate	oxybenzone	ethylene glycol
15	homosalate	propyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate
16	propyl 4-hydroxybenzoate	ethylene glycol	2,4-dihydroxybenzophenone
17	propyl 4-hydroxybenzoate	ethylene glycol	benzoic acid
18	propyl 4-hydroxybenzoate	ethylene glycol	butyl 4-hydroxybenzoate
19	propyl 4-hydroxybenzoate	520-36-5	methyl 4-hydroxybenzoate
20	ethylene glycol	methyl 4-hydroxybenzoate	styrene
21	propyl 4-hydroxybenzoate	ethylene glycol	styrene
22	oxybenzone	methyl 4-hydroxybenzoate	ethylparaben
23	homosalate	ethylene glycol	methyl 4-hydroxybenzoate
24	methyl 4-hydroxybenzoate	ethylparaben	2,4-dihydroxybenzophenone
25	methyl 4-hydroxybenzoate	ethylparaben	benzoic acid

The complete rankings are included in the Supplemental Material Individual Authoritative Lists. The individual chemicals ranked by EF and RF are in the EDCDB EF_C and EDCDB RF_C tables, respectively. The 2-, 3-, and 4-way combinations ranked by EF are in the EDCDB 2-way Combo (EF, Kantar), EDCDB 3-way Combo (EF, Kantar), and EDCDB 4-way Combo (EF, Kantar) tables, respectively. The 2-, 3-, and 4-way combinations ranked by RF are in the EDCDB 2-way Combo (RF, Kantar), EDCDB 3-way Combo (RF, Kantar), and EDCDB 4-way Combo (RF, Kantar) tables, respectively.

6.3.5 Ranking the Compounds from Dodson et al. (2012)

The EDC and asthma-associated chemicals selected by Dodson et al. (2012) are prevalent in consumer products, particularly among cosmetics, hair care, and personal care

products. Gabb and Blake (2016a) performed an informatics analysis on the same chemicals. These early results are summarized in Table 1 (the prevalence of each individual DODSON chemical), Table 7 (the prevalence of the DODSON chemicals by product category), and Figure 16 (the heatmap showing the percentage of each DODSON chemical by product category). The most common DODSON chemicals and product hotspots are readily apparent from this data. Phenoxyethanol (a glycol ether and common preservative) is the most frequently occurring chemical, followed by methyl paraben (another common preservative), the natural fragrances limonene and linalool, and octinoxate (a UV filter). These chemicals span many product categories. Cosmetics and hair care products have several hotspots for glycol ethers, fragrances, parabens, and to a lesser extent, UV filters. It is not surprising that UV filters are common in sunscreens and some cosmetics and hair care products. Personal care, hair care, and cosmetic products have hotspots for glycol ethers, natural fragrances, and parabens. "Fragrance" is the second most common ingredient in our product sample after water. Various flavors and flavorings also occur frequently. While the target chemicals limonene, linalool, and a few other natural fragrances are fairly common among products in our sample, the synthetic fragrance chemicals are comparatively rare.

Gabb and Blake (2016a) ranked the DODSON chemicals and their combinations by per-product prevalence across all consumer product categories. Figure 25, Table 33, and Table 34 show the per-consumer rankings of the DODSON chemicals in personal care products. Among the hotspots, fragrance products like perfumes have a lot of linalool and limonene, which is not surprising given that these are common fragrance chemicals. These chemicals are also common in hair styling products. The preservative, 2-phenoxyethanol, is common in makeup, moisturizers, shampoo, cleaners, and hair styling products. The paraben, methyl 4-hydroxybenzoate, is common in makeup and moisturizers. Among other highly ranked DODSON chemicals, parabens (methyl and butyl 4-hydroxybenzoate, and ethylparaben) and fragrance/flavor chemicals (linalool, limonene, lilial, and eugenol) appear in nearly every product category. The emollient, decamethylcyclopentasiloxane, is also common among personal care products. The UV filters, oxybenzone and 2,4-dihydroxybenzophenone, both have

high RF but the latter is only detected in four product categories. It is not surprising that the same chemicals in Figure 25 also comprise the 2- and 3-way combinations with the highest RF.

The DODSON heatmap below is sparser than those of TOX21 (Figure 21), HSDB (Figure 22), and CACOC (Figure 23). This is not an indication of the relatively danger or safety of the chemicals in this list. Rather, it indicates that DODSON does not provide the same level of coverage for the ingredient space of personal care products.

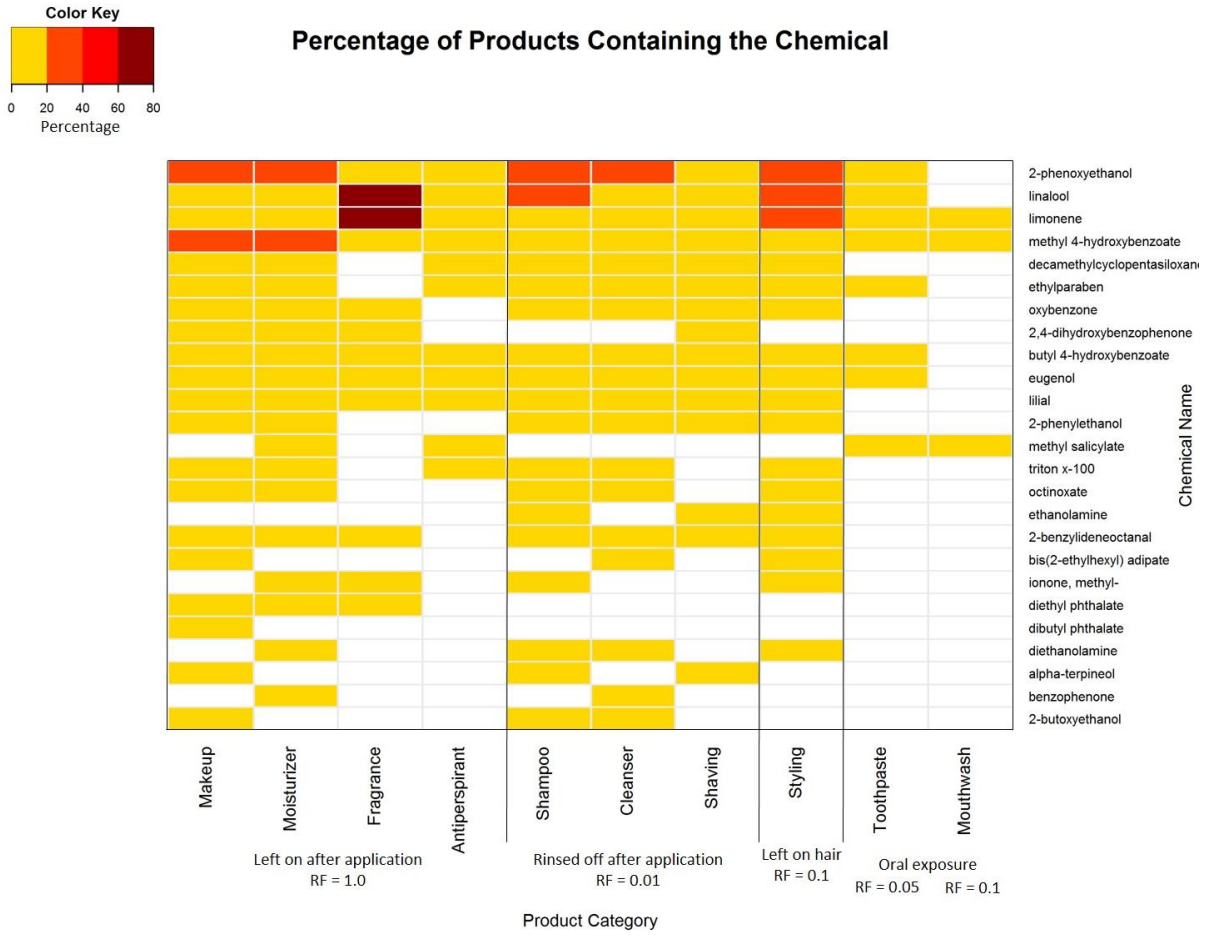


Figure 25 Heatmap of prevalence by product category for the top-25 DODSON chemicals

The chemicals are ranked top to bottom by RF score. White indicates that a chemical was not found in a product category. Yellow indicates that > 0 – 20% of the products in the category contain the chemical. Orange indicates that > 20 – 40% of the products contain the chemical. Red indicates that > 40 – 60% of the products contain the chemical. Dark red indicates that > 60 – 80% of the products contain the chemical.

Table 33 Top-25 2-way DODSON chemical combinations ranked by RF

	Chemical 1	**Chemical 2**
1	linalool	2-phenoxyethanol
2	limonene	2-phenoxyethanol
3	methyl 4-hydroxybenzoate	2-phenoxyethanol
4	linalool	limonene
5	methyl 4-hydroxybenzoate	linalool
6	methyl 4-hydroxybenzoate	limonene
7	decamethylcyclopentasiloxane	2-phenoxyethanol
8	ethylparaben	2-phenoxyethanol
9	decamethylcyclopentasiloxane	linalool
10	decamethylcyclopentasiloxane	limonene
11	ethylparaben	linalool
12	ethylparaben	limonene
13	oxybenzone	2-phenoxyethanol
14	2,4-dihydroxybenzophenone	2-phenoxyethanol
15	methyl 4-hydroxybenzoate	decamethylcyclopentasiloxane
16	2-phenoxyethanol	butyl 4-hydroxybenzoate
17	methyl 4-hydroxybenzoate	ethylparaben
18	oxybenzone	linalool
19	oxybenzone	limonene
20	2,4-dihydroxybenzophenone	linalool
21	eugenol	2-phenoxyethanol
22	2,4-dihydroxybenzophenone	limonene
23	linalool	butyl 4-hydroxybenzoate
24	limonene	butyl 4-hydroxybenzoate
25	oxybenzone	methyl 4-hydroxybenzoate

Table 34 Top-25 3-way DODSON chemical combinations ranked by RF

	Chemical 1	**Chemical 2**	**Chemical 3**
1	linalool	limonene	2-phenoxyethanol
2	methyl 4-hydroxybenzoate	linalool	2-phenoxyethanol
3	methyl 4-hydroxybenzoate	limonene	2-phenoxyethanol
4	methyl 4-hydroxybenzoate	linalool	limonene
5	decamethylcyclopentasiloxane	linalool	2-phenoxyethanol
6	decamethylcyclopentasiloxane	limonene	2-phenoxyethanol
7	ethylparaben	linalool	2-phenoxyethanol
8	ethylparaben	limonene	2-phenoxyethanol
9	methyl 4-hydroxybenzoate	decamethylcyclopentasiloxane	2-phenoxyethanol
10	decamethylcyclopentasiloxane	linalool	limonene
11	methyl 4-hydroxybenzoate	ethylparaben	2-phenoxyethanol
12	ethylparaben	linalool	limonene
13	oxybenzone	linalool	2-phenoxyethanol
14	oxybenzone	limonene	2-phenoxyethanol
15	2,4-dihydroxybenzophenone	linalool	2-phenoxyethanol
16	methyl 4-hydroxybenzoate	decamethylcyclopentasiloxane	linalool
17	2,4-dihydroxybenzophenone	limonene	2-phenoxyethanol
18	methyl 4-hydroxybenzoate	decamethylcyclopentasiloxane	limonene
19	linalool	2-phenoxyethanol	butyl 4-hydroxybenzoate
20	limonene	2-phenoxyethanol	butyl 4-hydroxybenzoate
21	methyl 4-hydroxybenzoate	ethylparaben	linalool
22	methyl 4-hydroxybenzoate	ethylparaben	limonene
23	oxybenzone	methyl 4-hydroxybenzoate	2-phenoxyethanol
24	methyl 4-hydroxybenzoate	2,4-dihydroxybenzophenone	2-phenoxyethanol
25	oxybenzone	linalool	limonene

The complete rankings are included in the Supplemental Material Individual Authoritative Lists. The individual chemicals ranked by EF and RF are in the DODSON EF_C and DODSON RF_C tables, respectively. The 2-, 3-, and 4-way combinations ranked by EF are in the DODSON 2-way Combo (EF, Kantar), DODSON 3-way Combo (EF, Kantar), and DODSON 4-way Combo (EF, Kantar) tables, respectively. The 2-, 3-, and 4-way combinations ranked by RF are in the DODSON 2-way Combo (RF, Kantar), DODSON 3-way Combo (RF, Kantar), and DODSON 4-way Combo (RF, Kantar) tables, respectively.

Chapter 7: Discussion

7.1 Assessing the Authoritative Lists of Potentially Harmful Chemicals

There is considerable overlap among the authoritative lists of potentially harmful chemicals (Figure 12). The overlap is not complete, however, so analyzing each list in isolation yields information about where the lists agree regarding chemicals that appear in consumer product formulations. Side-by-side comparison of the top-25 chemicals from each list reveals the degree of overlap with respect to RF (i.e., likely retention based on consumer usage patterns for personal care products) (Table 35). There is broad agreement that the parabens, methyl 4-hydroxybenzoate, propyl 4-hydroxybenzoate, and ethylparaben; the fragrance chemicals, linalool and limonene; the UV filter, oxybenzone (and to a lesser extent, 2,4-dihydroxybenzophenone and octinoxate); the cyclosiloxane, octamethyltrisiloxane (and to a lesser extent, decamethylcyclopentasiloxane); the glycol ether, 2-phenoxyethanol; the diol alcohol, ethylene glycol (and to a lesser extent, 1,2-propanediol); and acrylic acid are priority targets for individual risk assessment. Many other prevalent chemicals (e.g., 1-hexadecanol and 1-docosanol) are deemed priority targets by two out of five authoritative lists. The number of prevalent chemicals (e.g., acid blue 9 and methyl salicylate) that are deemed a priority target by only one list shows the diversity of the lists.

Table 35 Comparison of the top-25 chemicals from each authoritative list ranked by RF

	TOX21	HSDB	CACOC	EDCDB	DODSON
1	octamethyltrisiloxane	octamethyltrisiloxane	octamethyltrisiloxane	methyl 4-hydroxybenzoate	2-phenoxyethanol
2	glycerol	glycerol	silica	propyl 4-hydroxybenzoate	linalool
3	vitamin e acetate	1,2-propanediol	titanium dioxide	ethylene glycol	limonene
4	1,2-propanediol	butylated hydroxytoluene	methyl 4-hydroxybenzoate	ethylparaben	methyl 4-hydroxybenzoate
5	butylated hydroxytoluene	1-(2-butoxy-1-methylethoxy)propan-2-ol	propyl 4-hydroxybenzoate	oxybenzone	decamethylcyclopentasiloxane
6	silica	titanium dioxide	ethylene glycol	2,4-dihydroxybenzophenone	ethylparaben
7	1-(2-butoxy-1-methylethoxy)propan-2-ol	2-phenoxyethanol	acrylic acid	benzoic acid	oxybenzone
8	2-phenoxyethanol	linalool	sodium hydroxide	butyl 4-hydroxybenzoate	2,4-dihydroxybenzophenone
9	linalool	limonene	isobutane	styrene	butyl 4-hydroxybenzoate
10	limonene	raphisiderite	decamethylcyclopentasiloxane	homosalate	eugenol
11	methyl 4-hydroxybenzoate	methyl 4-hydroxybenzoate	isopropanol	520-36-5 (apigenin)	lilial
12	edta	alpha-tocopherol	ethylparaben	beta-carotene	2-phenylethanol
13	vitamin e	edta	retinol	caffeine	methyl salicylate
14	propyl 4-hydroxybenzoate	vitamin e	propane	2,2',4,4'-tetrahydroxybenzophenone	triton x-100
15	1-hexadecanol	propyl 4-hydroxybenzoate	oxybenzone	hydroquinone	octinoxate
16	1-docosanol	ethylene	aluminum	chrysin	ethanolamine
17	dl-panthenol	1-hexadecanol	butyl 4-hydroxybenzoate	2-methoxyethanol	2-benzylideneoctanal

Table 35 (cont.)

18	9011-80-7 (adipic acid, phthalic anhydride, dipropylene glycol resin)	1-docosanol	styrene	hydrogen sulfide	bis(2-ethylhexyl) adipate
19	ethylene glycol	acid blue 9	acrylamide	octinoxate	ionone, methyl-
20	acrylic acid	dl-panthenol	acid violet 49	resorcinol	diethyl phthalate
21	tween 20	9011-80-7 (adipic acid, phthalic anhydride, dipropylene glycol resin)	acid red 87	4-chloro-3,5-dimethylphenol	dibutyl phthalate
22	1,3-butanediol	ethylene glycol	methyl methacrylate	4-hexylresorcinol	diethanolamine
23	hexitol	acrylic acid	phosphoric acid	ferulic acid	alpha-terpineol
24	potassium sorbate	tween 20	silver	toluene	benzophenone
25	benzyl alcohol	1,3-butanediol	propene	phenol	2-butoxyethanol

Colors indicate the degree of agreement among the lists. Red indicates chemicals that are high priority in four or five lists. Orange indicates chemicals that are high priority in three lists. Yellow indicates chemicals that are high priority in two lists.

7.2 Limitations of the Informatics Approach

Like any experimental method, the informatics approach is not without limitations. First, information for a large sample of consumer products must be readily available. Only products in the Drugstore.com inventory were analyzed in the present study. This does not represent every consumer product currently on the market but it provides a reasonable cross-section of general consumer products (Table 7), personal care products (Table 8), and most importantly, their formulations. Second, the chemical dictionary must provide adequate coverage of the chemical namespace. PubChem, the dictionary used to map chemical names to unique identifiers, provides excellent (Table 10 and Supplemental Material, Target Chemicals) but incomplete (Table 13 and Supplemental Material, Unmatched Ingredient Analysis) coverage of the target chemicals. Third, only products that provide an ingredient list can be analyzed using the informatics approach. Likewise, only chemicals that are actually listed can be detected. GCMS analyses detect potentially harmful chemicals that do not appear in ingredient lists (Dodson et al., 2012; Steinemann et al., 2011; Steinemann, 2015). For example, chemicals that are leached from product packaging will not appear in an ingredient list because they are not part of the product formulation. However, the informatics approach can analyze a much larger product sample than spectrographic approaches. The larger sample provides a more comprehensive view of product formulations, and hence can detect chemicals that are missed by GCMS (Gabb and Blake, 2016a). Therefore, informatics and GCMS are complementary approaches.

Current product labeling regulations in the United States do not require manufacturers to disclose trade secrets. Fragrance and flavor mixtures are often treated as trade secrets, so the individual chemicals in such mixtures are not disclosed. They are simply listed as generic “fragrance” or “flavor” on product labels, which can mask the presence of many chemicals (Steinemann et al., 2011; Steinemann, 2015). Steinemann’s (2015) GCMS analysis detected 156 volatile chemicals in 37 consumer products, many of which are fragrances. Many of the detected chemicals are not consistently listed on product labels. However, not all of the 11,964 target chemicals in TOX21, HSDB, CACOC, EDCDB, and DODSON are volatile fragrance compounds. Also, many fragrance chemicals do commonly appear on product labels (e.g., limonene, linalool, benzyl alcohol), though their frequencies may be underestimated

(Steinemann, 2016; Gabb and Blake, 2016b). Once again, informatics and GCMS should be considered complementary approaches. Fortunately, legislative pressure and industry trends are moving toward greater transparency (Service, 2009; Schmidt, 2016; Nicole, 2018).

Third, EF and RF were only computed for the subset of product categories in the Kantar dataset (Table 8). Ideally, a more comprehensive, though more expensive, dataset of consumer usage patterns could be obtained that would allow the prevalent chemical combinations in the entire product sample to be determined. In the meantime, the categories in the available Kantar dataset represent personal care products that might typically be used daily for long periods of time, as opposed to occasional-use products like cold medicine, medicinal ointments, teeth whiteners, wart removers, or insecticides.

7.3 Evaluating the Ranked Lists without *A Priori* Knowledge of Biological Activity

The target chemicals were drawn from five authoritative toxicology lists: TOX21, HSDB, CACOC, EDCDB, and DODSON. There is considerable overlap among these lists (Figure 12 and Supplemental Material, Target Chemical Overlap). Fifteen chemicals appear in all five lists, but this should not be interpreted as widespread agreement that a chemical is harmful. Most of the target chemicals (8,622 out of 11,964 distinct chemicals) appear in only one list, which indicates that these lists have different inclusion criteria. Therefore, membership in these lists could mean a number of things. For example, toxicological evidence may suggest that the chemical is harmful, or at least potentially harmful under certain circumstances. It could also mean that the probability of environmental exposure is sufficiently high (e.g., due to high production volume) that a regulatory agency deems the chemical a research priority.

Whatever the reason for inclusion in one or more of the authoritative lists, the present study avoids value judgments (i.e., assumptions about relative safety and harm) about the target chemicals. However, as some lists contain relatively benign chemicals (e.g., water, glycerol, sodium chloride, sucrose, etc.), some thought was given to filtering endogenous chemicals from the list of targets. This idea was rejected for four related reasons. First, categorization is not straightforward because the definition of endogenous and exogenous chemicals is ambiguous. For example, the human body does not produce vitamin c (which is why we are susceptible to scurvy), but it is typically present in a healthy person. Is it

endogenous because it is normally present in the body, or is it exogenous because it enters the body externally? Second, an objective method of filtering endogenous chemicals could not be found. Third, endogenous versus exogenous categorization is fraught with value judgments about the chemicals. For example, vitamin c is generally considered necessary and beneficial, but the possibility exists that it could be harmful in combination with other chemicals. Fourth, treating the target chemicals equally did not produce results that were dominated by seemingly unimportant chemicals. For example, glycerol, sodium chloride, and vitamin c did not drown the signal from other chemicals. The question of distinguishing endogenous and exogenous chemicals should be revisited in future work because it is an important consideration for environmental chemists (Dennis et al., 2017, p. 509):

> *"Develop chemistry methods to enable the detection of low-abundance chemicals and to enable differentiation of endogenous molecules from exogenous molecules. Through methods such as multiplexing, interfering chemicals can be removed to allow detection of low-level environmental chemicals that are often difficult to detect because of higher-abundance endogenous chemicals from food, drugs, and normal metabolic processes (Rappaport et al., 2014). Investments in the development of semi-targeting or multiplexing strategies should be a high priority."*

In the end, only water (which appears in the HSDB list) was removed from the list of target chemicals. Though water intoxication, an acute form of poisoning, can occur, humans are an aqueous medium and water is essential to biochemical processes. Water is also the main solvent in most liquid consumer products, so including it would drown the signal from the other target chemicals.

The goal of the present research is to objectively prioritize the target chemicals based on near-field exposure from consumer products. Evaluating the quality of the prioritization is far more difficult than evaluating individual system components. There are several objective ways to prioritize the target chemicals for in-depth toxicological testing. For example, a chemical's production volume can be used as a proxy for far- and near-field exposure (Muir and Howard, 2006; Sanderson et al., 2006, 2013). The rationale is that HPV chemicals will eventually find their way into the environment due to inadequate sequestration (e.g., chemical waste from industrial processes), direct release (e.g., pesticides used in industrial farming), or use in

consumer products. The Chemical Data Reporting Rule of the Toxic Substances Control Act requires manufacturers and importers to provide the EPA with data on the chemicals they produce or import into the U.S. A prioritized list of HPV chemicals could be derived from this data (http://www.epa.gov/chemical-data-reporting). Structure-activity relation modeling is another approach to prioritization (Dellarco et al., 2010; Tice et al., 2013; Wang et al., 2011, 2012). The rationale is that structurally similar chemicals will have similar biological activity, so chemicals with unknown toxicity are mapped to those with known toxicity and ranked accordingly. High-throughput *in vitro* assays are yet another approach to prioritization (Huang et al., 2016), the assumption being that *in vitro* toxicity accurately predicts human toxicity.

All of these approaches are scientifically valid but evaluating the resulting prioritizations is a hard problem because the relative danger of the target chemicals is not known *a priori*. It is possible that the highest ranked chemical in any of these schemes is harmless at typical human exposure levels. The ranked lists could be shown to toxicologists to see if they conform to expert expectations. This is unsatisfying because if intuition was sufficient, objective ranking would not be necessary. Also, no toxicologist is an expert on all chemical classes so bias toward a particular class or subset of classes is likely. The prioritized lists from several approaches could be compared for commonalities among the top-ranked chemicals, but this is also unsatisfying because commonalities may just be coincidence. For example, the chemicals with the highest *in vitro* toxicity scores may not even be present in consumer products, or have very low environmental exposure levels. The relative effectiveness of prioritization schemes will only be known after the target chemicals are subjected to in-depth toxicological testing, and perhaps not even then, because the previous discussion is predicated on notion that the best prioritization scheme will rank harmful chemicals at the top of the list. For the sake of argument, let us say that *post facto* evidence indicates the top-ranked chemicals (regardless of prioritization approach) are harmful. Is this result necessarily better than ranking harmless chemicals at the top of the list? It is just as important to know that a chemical is safe, especially if its likely exposure is high and its alternative, if one exists, is significantly more expensive. The EPA NexGen framework for evidence-based population health risk management must take competing factors into account (Krewski et al., 2014). It is expected to take a long time but the

goal of the Tox21 consortium is to assess every chemical in the Tox21 10K Library. The argument of the present research is that near-field exposure from consumer products is an intelligent way to prioritize chemicals for risk assessment because it accounts for potential impact to the broad population.

As noted above, there are many scientifically valid approaches to prioritizing chemicals for CRA (e.g., by production volume, high-throughput *in vitro* screening, SAR modeling, and exposure/retention modeling). Though it is not currently possible to conclude that one ranking is *better* than another, it can be argued that accounting for consumer usage patterns and relative chemical retention leads to a more *rational* ranking.

Chapter 8: Conclusions and Future Work

8.1 Effectiveness of Informatics Approaches to Near-Field Chemical Exposure

The present research applies an informatics approach to the analysis of potentially harmful chemicals in everyday consumer products. It extends the preliminary analysis of Gabb and Blake (2016a) using a small set of 55 chemicals from a recent GCMS analysis (Dodson et al., 2012). These are the DODSON chemicals, which are suspected endocrine disrupting and/or asthma-associated compounds. They were found to be common among the 38,975 products in the original CPDB (Table 1, Table 7, and Figure 16) – further evidence that consumer products contribute to near-field exposure.

One advantage of an informatics approach is the number of target chemicals and products that can be considered. The cost and labor involved in GCMS make it impractical to analyze tens of thousands of target chemicals among tens of thousands of consumer products. The GCMS analysis only tested 213 different products in 42 composite samples (Dodson et al., 2012). The informatics approach found products with target chemicals that are not detected in the small GCMS sample. For example, the informatics approach showed that toothpastes contain the same three target chemicals found in the GCMS analysis: the antimicrobial triclosan and the natural fragrances methyl salicylate and eugenol. However, several more of the target chemicals also appear in toothpaste formulations: phenoxyethanol, linalool, limonene, butyl paraben, ethyl paraben, and methyl paraben (Figure 16). The antimicrobials further demonstrate the utility of the informatics approach. Triclocarban was detected in four product categories (bar soap, facial cleanser, liquid soap, and deodorant and antiperspirant) (Figure 16), whereas it was only detected in one GCMS sample (bar soap). The CPDB contains triclosan in 17 product categories (Figure 16), compared to only three of the GCMS samples. Finally, Dodson et al. (2012) only analyzed six product categories for UV filters (sunscreen and shaving cream) and cyclosiloxanes (sunscreen and car interior cleaners). By comparison, the CPDB contains UV filters and cyclosiloxanes in 22 product categories (Figure 16).

However, the informatics approach also has limitations relative to GCMS, as noted in Chapter 7.2. In a nutshell, the informatics approach can only detect chemicals that appear in a

product ingredient list. It will not detect chemicals leached from product packaging, degradation byproducts, or chemicals in propriety mixtures that are not disclosed.

The informatics approach reveals the degree to which chemical synonymy undermines the informed consent that product labeling regulations are meant to provide. Unless steps are taken to reduce obfuscation and improve chemical literacy, chemical synonymy can hinder consumer decision-making with respect to the chemicals in their products. For example, suppose that consumers trying to manage their asthma read a news article claiming that a specific fragrance chemical may exacerbate asthma attacks. They check the ingredient lists on the products in their homes and feel satisfied that none of them contain the fragrance. This is a false sense of security unless they have also checked for commonly used synonyms for the fragrance that may not have been mentioned in the news source. This same scenario can be applied to many other chemical ingredients, as illustrated in Table 1. Apply the reverse logic to a consumer looking for a fragrance-free product. Many products only specify “fragrance” (the second most common ingredient after water) on the ingredient label instead of listing each fragrance chemical in the mixture. These products are easy to avoid. Ironically, products that explicitly list fragrance chemicals may be harder for a consumer to assess. Consider a product that lists butylphenyl methylpropional but not “fragrance” in the ingredient label. Unless the consumer knows that this is a fragrance chemical, he may mistakenly assume that the product is fragrance-free. Risk perception adds another dimension to the problem of chemical synonymy. Namely, consumers may choose a product that lists wintergreen oil as an ingredient instead of one that lists methyl salicylate because the product with wintergreen oil seems more “natural,” in spite of the fact that they denote the same chemical.

The primary goal of the informatics approach, however, is to inform decisions about which chemicals and chemical combinations to subject to toxicological analysis and CRA. Few of the 80,000+ chemicals registered under the U.S. Toxic Substances Control Act of 1976 have received much study (Judson et al., 2009; IOM, 2014), and even fewer have been subjected to CRA. Many of these chemicals are used in consumer product formulations, and a typical consumer will use these products daily for long periods of time. Given that so few chemicals have undergone toxicological testing, and that their individual and combined biological

activities are largely unknown, the safety of this persistent, near-field exposure should not be assumed (Boekelheide and Campion, 2010; Dennis et al., 2016). CRA considers multiple stressors but performing risk assessment on all possible chemical mixtures is infeasible. It is expensive and time-consuming, so it is unlikely that each individual chemical will be tested much less all of their possible combinations. Therefore, intelligent prioritization is required.

There are many ways to objectively conduct the necessary prioritization. For example, one could assume that high production volume eventually leads to far-field exposure because such chemicals are prevalent in the environment and persist for long periods of time. Given a set of chemicals with known biological activity, it is possible to group chemicals with unknown activity based on structural similarity (i.e., SAR modeling). HTS can rapidly test the *in vitro* biological activity of individual chemicals and chemical combinations. These approaches have advantages and disadvantages, and in the absence of a priori knowledge of harm or safety, they all have some degree of validity. However, with the exception of HTS, these prioritization approaches say little about combined chemical exposure. The informatics approach described here prioritizes testing based on near-field exposure from everyday consumer products. The selected target chemicals and their combinations were prioritized based on their prevalence within individual products, and their likely exposure and retention based on consumer usage patterns. This prioritization approach scales to tens of thousands of target chemicals and tens of thousands of consumer products. Potential improvements to this approach are proposed in Chapter 8.2.

"Computational exposure science represents a frontier of environmental science that is emerging and quickly evolving," according to a review of the field by Egeghy et al. (2016, p. 697). These authors go on to say that "computational exposure science, linked with comparable efforts in toxicology, is ushering in a new era of risk assessment that greatly expands our ability to evaluate chemical safety and sustainability and to protect public health." (p. 697) The present research aims to advance the relatively new field of personal chemical exposure informatics, which combines exposure-based chemical prioritization, consumer exposure models, chemical information for consumer products, exposure factors and informatics, participatory methods and personal informatics, and community engagement (Goldsmith et al.,

2012). The informatics approach described here addresses aspects of the first four directions but not the last two, though future work could encompass these areas (Figure 26).

8.2 Future Work

8.2.1 Incorporating HTS Data from ToxCast

The target chemicals used in the present study were drawn from five authoritative lists of potentially harmful chemicals: TOX21, HSDB, CACOC, EDCDB, and DODSON. As noted previously, some of the chemicals in these lists are relatively benign (e.g., vitamin c, glycerol, sucrose, sodium chloride) compared to others (e.g., triclosan and triclocarban, which were recently banned from certain consumer products). For example, glucose and caffeine are among the target chemicals; both are known endocrine "disruptors" with widespread exposure in the worldwide population. However, glucose is the critical component of several metabolic pathways, and is naturally present in healthy people. Likewise, caffeine has been consumed in high doses for centuries, perhaps even millennia. This does not necessarily make them safe, because both chemicals are associated with disease endpoints: glucose affects diabetes and caffeine affects hypertension. Are glucose and caffeine safe or harmful? The answer to this question is ambiguous, so filtering "safe" chemicals from the list of targets is fraught with subjective value judgments, as discussed in Chapter 7.3. For this reason, the present study takes an entirely value-neutral view of the target chemicals. The authoritative lists were compiled by toxicologists, so they were taken to be just that – authoritative.

ToxCast offers a promising future direction to improve the signal-to-noise ratio of chemicals with greater potential for harm (Dix et al., 2007; Richard et al., 2016). ToxCast measures the *in vitro* activity of 9,076 chemicals using 1,192 HTS assays (EPA, 2015b). In theory, these results can be used to filter chemicals that are generally regarded as safe while giving greater weight to those deemed harmful. The informatics approach would then be prioritizing chemicals and chemical combinations for CRA based not only on likely exposure and retention but also on biological activity.

However, this theory is controversial for a number of reasons. First and foremost, *in vitro* assays do not necessarily predict *in vivo* behavior. It is not yet known how well the ToxCast assays model real biological endpoints. Their reliability has been called into question (Janesick

et al., 2016; Houck et al., 2017; Janesick et al., 2017). Second, biological activity does not necessarily equal harm. Third, the HTS approach used by ToxCast is adapted from pharmacology (Janesick et al., 2016). Toxicology and pharmacology may be two sides of the same coin, but they have very different goals (Janesick et al., 2016, p. 1225):

> *"Another issue is that the assays used in ToxCast were largely pre-existing commercial assays which were adopted from the philosophy and approach of the pharmaceutical industry. Assays for drug discovery are designed to identify only the strongest hits in large libraries of structurally similar chemicals (millions or more) to limit the subsequent screening required to develop lead compounds for preclinical studies. This is philosophically the opposite of a proper chemical genomics approach to identify potential bad actors that should be selected for further scrutiny. Such assays would seek to identify every chemical that activates a particular pathway in a statistically significant way and then rank these for further testing. The ability of ToxCast assays to predict* in vivo *toxicity is often evaluated by comparing the effects of a chemical in ToxCast with effects from guideline studies*, in vivo *(Rotroff et al., 2013). However, the end points in guideline studies are not always sensitive to chemical effects on the endocrine system (Zoeller et al., 2012); thus, limiting their utility as validators of ToxCast assays for endocrine activity."*

Many environmental toxicants (particularly EDCs) act at very low doses (like most human hormones), often below accepted no-observed-effect-levels, with downstream health consequences that are only apparent years after exposure (Grun and Blumberg, 2006; Janesick and Blumberg, 2011). Also, in some pharmacological assays (e.g., testing potential chemotherapy drugs on tumor cells), cell death could be the desired result, whereas toxicological assays try to measure subtle cellular changes without actually killing the cells or stressing them too severely. Therefore, ToxCast includes 35 "burst assays" to help filter chemicals that are simply too toxic for HTS (Judson et al., 2016, p. 324):

> *"Many chemicals show activation of large numbers of assays ... in which cell stress and cytotoxicity are also seen. We term this phenomenon the cytotoxicity-associated 'burst'... In such situations, activity represents a false positive..."*

Fourth, the ToxCast results for a given chemical are subject to interpretation. For example, the chemotherapy drug, tamoxifen citrate, has 241 positive and 645 negative assays. It is also positive for 34 out of 35 burst assays. Is it active, inactive, or too cytotoxic to assess with HTS?

Tamoxifen is known to be biologically active, which is why it is an effective breast cancer treatment, but the large number of positive burst assays indicate that it may be too cytotoxic for the subtle responses that ToxCast is designed to test. Tamoxifen is not an ingredient in everyday consumer products, but it illustrates the ambiguity of some ToxCast results.

The ToxCast assay results for the target chemicals detected in the CPDB are tabulated in Supplemental Material CPDB ToxCast Assay Summary. Among these chemicals, three examples show how ToxCast could be used to change the current value-neutral approach to prioritization. Rhodamine 6g (a fluorescent dye) has 13 positive burst assays (out of 18 tested), so its 95 positive regular assays (out of 178 tested) may be false positives. The positive assays could point to a toxicological endpoint, or the chemical could just be poisoning the cell cultures. Should rhodamine 6g be given more weight during prioritization because the large number of positive assays (including burst assays) indicates biological activity? Unfortunately, there is no established threshold on how many positive burst activities indicate that the regular assays are suspect, nor is there a threshold on how many positive regular assays are needed to indicate biological activity. These thresholds would have to be set subjectively before deciding whether to give rhodamine 6g more weight during prioritization.

Some chemicals have less ambiguity. Ethoxyquin (a preservative) has many active regular assays (109 out of 457 tested) and no active burst assays (0 out of 35 tested), so it should have more weight during prioritization. A constant weighting factor could be used for all active chemicals or the ratio of active to inactive assays could be used as the weighting factor. However, this is a subjective decision.

On the other hand, it should be relatively easy to filter compounds that are generally regarded as safe. For example, sucrose is only active in two assays (out of 539 tested) with no cytotoxic activity. Citric acid, glucose, glycerol, and various other benign compounds have similar profiles. Filtering such chemicals from the targets is a straightforward, data-driven decision that should improve the signal from chemicals that are more likely to be harmful.

There is no single metric in the ToxCast data that can be used to weight the chemicals for potential harm. The ToxCast dataset (EPA, 2015b) is extensive and well-documented, but it is still a work-in-progress. The data are messy and often ambiguous. Therefore, careful analysis

is required before integrating ToxCast data into the prioritization scheme (Houck et al., 2017, p. A9):

> *"An important lesson of the ToxCast program is that no individual assay or data point should be considered in isolation or taken as 'truth' without consideration of the broader assay and data context, a host of technical and statistical factors derived from experience in working with the data, and both potency and efficacy."*

8.2.2 Using Chemical Absorption Models When Computing RF

The RF computations described in Chapter 5 take the product usage mode into account during prioritization. Chemical ingredients in products that are left on after application have higher weight during prioritization than those in products that are rinsed off after application (SCCS, 2015). The rationale is that exposure duration directly affects how much of a chemical is absorbed by the body. Though important, exposure duration is not the only factor affecting absorption. Some chemicals are absorbed more readily than others. Lipophilic chemicals (i.e., those that dissolve in fats and oils) pass through the dermis more easily than hydrophilic chemicals (i.e., those that dissolve in water). However, lipophilicity is hard to quantify, so a chemical's dermal permeability coefficient cannot be computed from first principles. Also, some chemicals are metabolized within the skin, which affects final absorption (Anissimov et al., 2013; Sugino et al., 2017). Absorption models exist for a few chemicals (e.g., Banyiova et al., 2016; Frederiksen et al., 2016; Sugino et al., 2017), but certainly not all of the target chemicals examined in this study. However, it might be possible to generalize these models and apply them to chemicals with similar characteristics (Anissimov et al., 2013; Frasch and Barbero, 2013; Hansen et al., 2013). If not, generalized dermal/oral absorption kinetics (Hall et al., 2007, 2011; Loretz et al., 2005, 2006, 2008; McNamara et al., 2007) can be used (Comiskey et al., 2015).

8.2.3 Taking Advantage of Unused Product Data

More data was scraped from the Drugstore.com product webpages than was actually used. In addition to product brand, name, ingredients, and retail hierarchy information that was so vital to this study, the CPDB also contains the following unused data: product price per unit weight or volume; textual product descriptions, usage instructions, and warnings; active

ingredients and their concentrations; and ingredient order. These data could be used to refine existing results or to begin new lines of inquiry.

They could also help answer sociotechnical questions not considered in the present analysis. For example, are cheaper consumer products more likely to contain the target chemicals? It may be possible to answer these questions using the price data in the CPDB. Do products marketed as green (or all-natural) really contain fewer of the target chemicals? GCMS analysis of a small product sample suggests that the chemical makeup of green and non-green products is largely the same (Dodson et al., 2012). Natural-sounding names are often used in place of chemical names (Gabb and Blake, 2016a). For example, lemon oil and wintergreen oil are synonyms of 1-methyl-4-prop-1-en-2-ylcyclohexene and methyl 2-hydroxybenzoate, respectively, both of which are among the target chemicals. If supposedly green products can be distinguished from ordinary products, it may be possible to answer questions about formulations in these product groups. A product's brand, name, and location in the retail hierarchy often indicate whether it is marketed as green. Textual product descriptions in the CPDB that are not currently used could also indicate marketing intent. Text mining techniques could reveal which products are marketed as green. Similarly, many products are aimed at specific consumer groups (e.g., children, the elderly, men, women, expectant/lactating mothers, or ethnic minorities). If products can be distinguished by consumer group, it may be possible to answer questions about relative chemical composition. Once again, mining the textual data may be the key. Finally, the additional textual data could be used to augment the current approach to assigning product categories using the retail hierarchy (Chapter 4.1.2).

The FPLA (§ 1454.c.3.B) requires manufacturers to list ingredients in "descending order of predominance" and to disclose active ingredients and their concentrations. Active ingredients are known to be biologically active, but in the present analysis they are not treated differently from other ingredients in the product formulation. Perhaps they should be given more weight (as a function of their concentrations) during prioritization. Similarly, position in the ingredient list, which contains some information about relative concentrations, could also be used to weight chemicals during prioritization. Isaacs et al. (2016, 2018) recently developed a probabilistic model to estimate the weight fraction of ingredients based on their position in

the ingredient list and the product category. Their model could be included as an additional weighting factor during prioritization.

8.2.4 Reengineering the Informatics Workflow

The consumer product and chemical landscape is constantly changing. Product formulations change as chemical prices fluctuate or new, cheaper or more effective chemical ingredients become available. Retail inventories also change with market conditions or as old products are discontinued and new products are brought to market based on consumer preferences. PubChem is also updated frequently with new chemicals and chemical synonyms. The current informatics workflow was designed to take snapshots of the consumer product landscape in order to answer the research questions about chemical prevalence. It was not designed to study the changing consumer product landscape. Also, like most research software, it "...is written to be good enough for the job intended" (Barnes, 2010). Version control, unit testing, long-term maintenance, extensibility, end-user experience, and other software engineering principles were secondary concerns. However, "an immediate challenge in computational exposure science is identifying and integrating data streams..." (Egeghy et al., 2016, p. 699).

Reengineering the informatics workflow into a continuous, event-driven system would keep the chemical prevalence and prioritization results always up-to-date with the latest data. Such a system would also make time series analysis of product formulations easier. Adding a user interface to the processed data would make the research results accessible to a broader audience rather than just those who are able to run the collection of scripts and modify the input data when necessary. A cloud-based system that includes continuous, event-based data updates and a user interface is proposed below (Figure 27). Implementation is beyond the scope of this research, but the proposed system illustrates several key capabilities that would transition the current ad hoc, proof-of-concept system to a true production system. Egeghy et al. (2016) provide a framework for computational exposure science. The current informatics workflow already covers several portions of their framework (Figure 26). The CPDB contains product formulations and could be used to study changes over time. The Kantar consumer product usage patterns provide information about product use. EF and RF provide information

about exposure and dose, respectively. Future research and the production system described in Figure 27 have the potential to cover several more. The relationship between product formulation and product category is likely to provide information about the functional role of some ingredients. Citizen science and sociotechnical approaches could yield information about consumer purchasing decisions. Media concentration could be covered by incorporating weight fraction models (Isaacs et al., 2016, 2018).

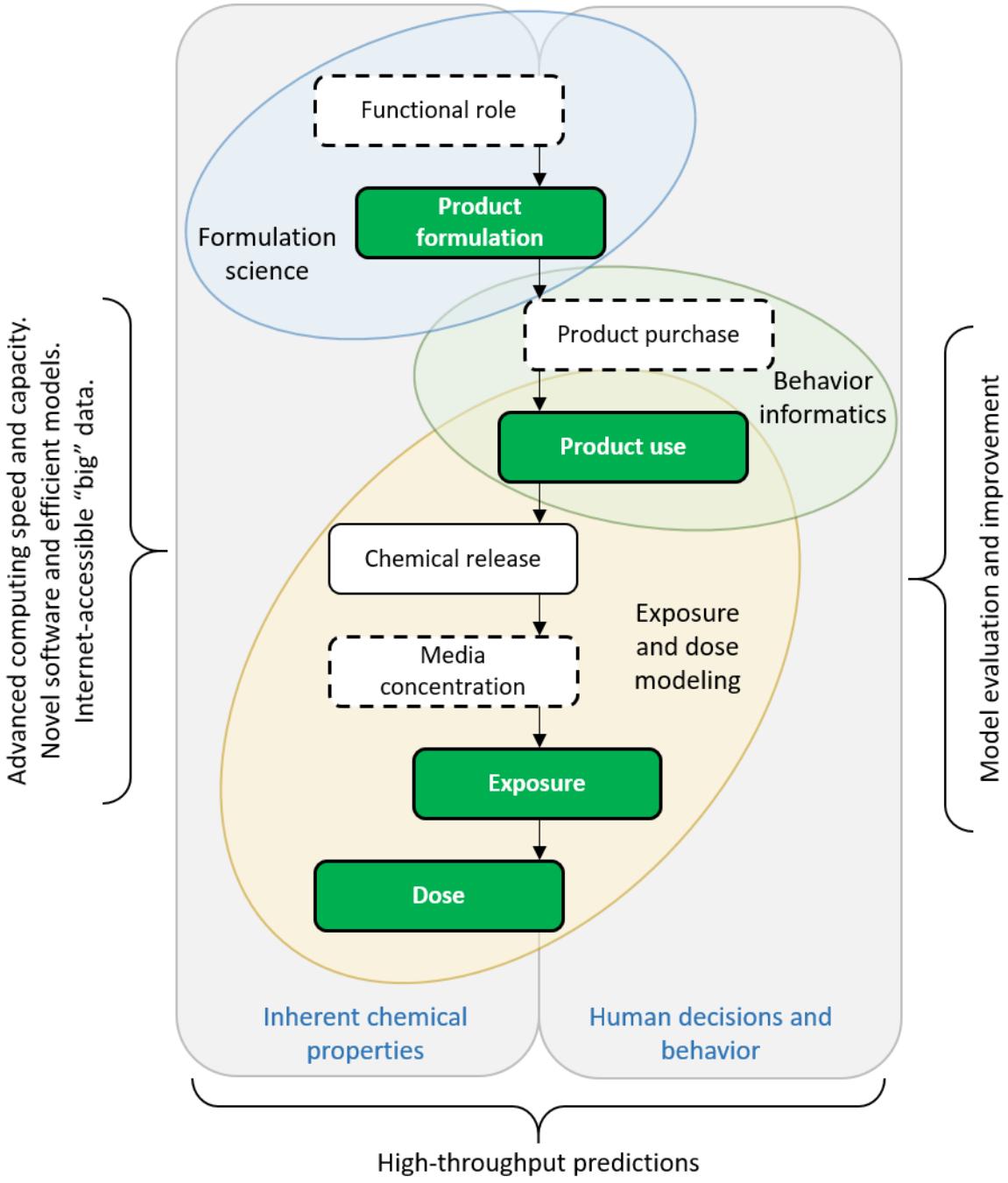


Figure 26 Computational exposure science framework

This framework (adapted from Egeghy et al., 2016) shows the interplay between consumer product chemistry, consumer behavior, and computational modeling. It highlights the key components of computational exposure science. The green rounded rectangles indicate components that the present informatics approach already covers. Rounded rectangles with dashed outlines indicate components that could be covered in future research.

The proposed system is based on the Amazon Web Services (AWS) platform, but other cloud service providers (e.g., Google Cloud Platform, Microsoft Azure, or IBM Cloud) offer similar capabilities. It relies on five AWS services: Elastic Compute Cloud (EC2), Elastic File

System (EFS), Simple Storage System (S3), Lambda, and Aurora. EC2 provides a way to launch virtual machines (VM) that mimic the computing environment used to run the workflows described in Chapter 6. The VMs can be sized to the task at hand. Simple tasks can be executed on lightweight, less-expensive VMs (i.e., those with fewer low-speed compute cores and smaller memory) while compute- or memory-intensive tasks are executed on more capable VMs. Parsing the large PubChem files or performing the combinatorial analysis would fall into the latter category. Some small tasks will be undertaken via the Lambda service, which is discussed in more detail below. EFS provides scalable and persistent file storage that can be mounted by the VMs running within EC2. Many of the scripts and intermediate data files described in Chapter 6 will reside on EFS for ready access, while the cheaper S3 object store will be used for bulk storage of raw HTML product pages and PubChem downloads. Though not shown in Figure 27, old or rarely-used data can eventually be moved to even cheaper archival storage (i.e., the AWS Glacier service). Much of the processed, structured data (e.g., product brand, name, category, and ingredient information; PubChem synonyms; target chemicals; Kantar Worldpanel dataset; heatmap; and chemical prevalence and combinatorial data) will be loaded into the Aurora relational database for easy querying by end-users.

Several aspects of the proposed system (Figure 27) require further explanation. The goal is to automate the workflows described previously in order to provide the most up-to-date data. It begins with periodic checks for new data. A scheduled task within an EC2 VM will periodically check the NCBI for PubChem updates. This can be accomplished using the standard Linux `cron` utility. If a newer version is available, the VM will download and parse it. The original file will be stored in S3 and the new chemicals and synonyms will be loaded into the Aurora database. To update product information, scheduled processes in Lambda will periodically retrieve the sitemaps from online consumer product retailers that allow web scraping. Lambda is a function-as-a-service platform to run small, short-lived functions inexpensively without the overhead of launching a VM. Whenever a sitemap is retrieved, it will be parsed to find new products. Their URLs will be stored in S3, which provides an event signaling mechanism to trigger processing whenever new data arrive. In the proposed system, the arrival of new product URLs in S3 triggers web scraping functions registered in Lambda to

retrieve the product pages, extract the required information, and update the consumer product database in Aurora. The arrival of new PubChem or product data in Aurora triggers the informatics workflow to map the target chemicals and product ingredients to unique identifiers, compute prevalence within the product sample, and reprioritize the chemicals and chemical combinations based on EF and RF.

There exists the possibility that individuals concerned about chemical exposure might want to access the results of this research or even contribute to it. Therefore, the proposed system (Figure 27) features two new capabilities: direct querying by end-users and direct updates by citizen scientists. The former capability is simply a matter of opening certain database tables in Aurora to SQL select queries or providing an interface to a set of predefined queries. Implementing the latter capability is more complicated because the integrity of the database is at stake. A system is envisioned whereby altruistic citizen scientists (Hand, 2010) can add consumer products that are inaccessible to web scraping and validate the parsing of product webpages, but the database must be protected against accidental damage by well-meaning amateurs and deliberate damage by vandals. This is a well-known problem on Wikipedia, for example, and much work goes into detecting digital vandalism (Geiger and Ribes, 2010; Potthast, 2010). However, a combination of manual screening by the system administrators and cross-validation by contributors should mitigate these risks.

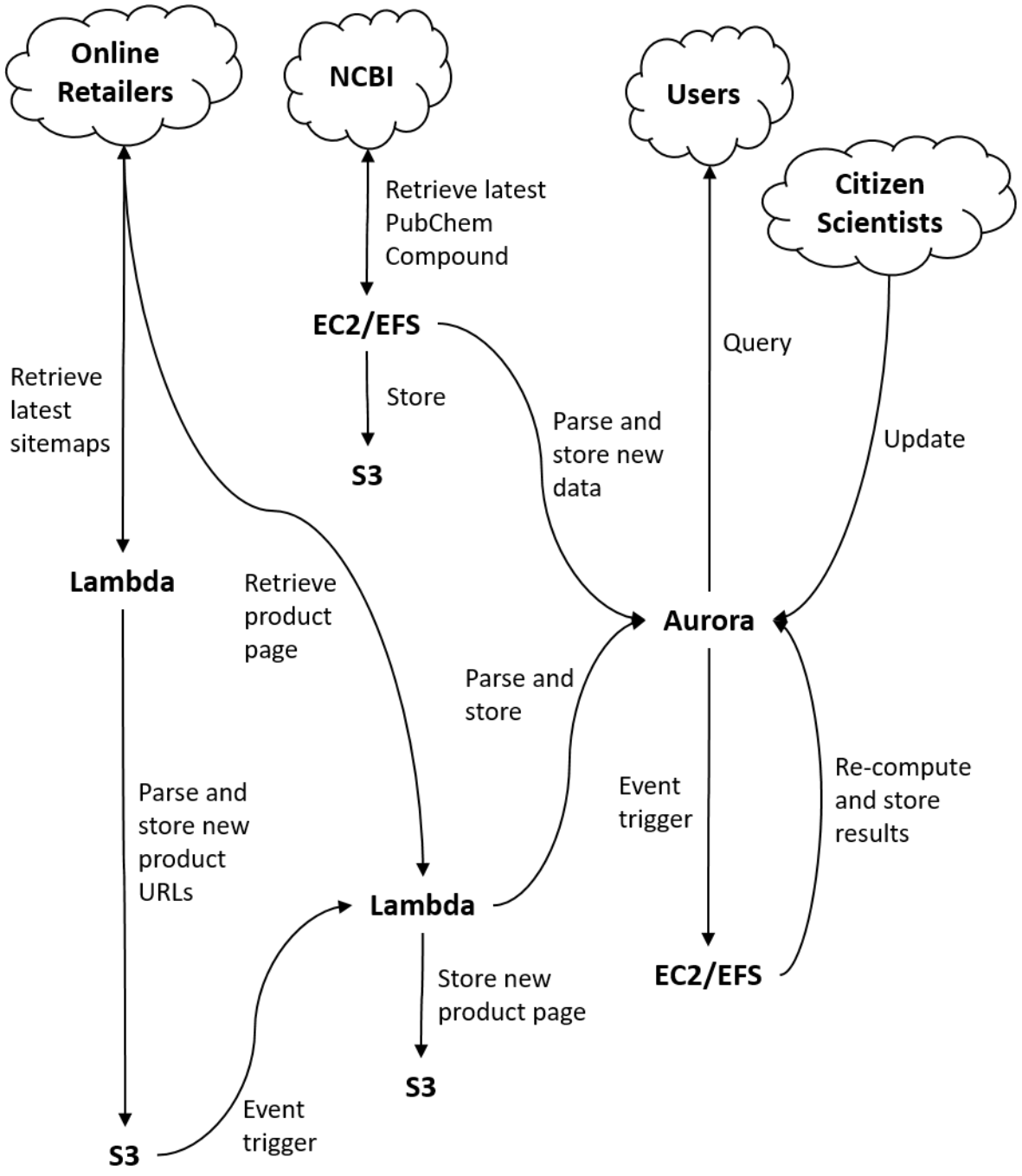


Figure 27 Event-driven, asynchronous system to prioritize the target chemicals.

References

Anderson, S. E., Franko, J., Kashon, M. L., Anderson, K. L., Hubbs, A. F., Lukomska, E., & Meade, B. J. (2013). Exposure to triclosan augments the allergic response to ovalbumin in a mouse model of asthma. *Toxicol Sci, 132*(1), 96-106.

Anissimov, Y. G., Jepps, O. G., Dancik, Y., & Roberts, M. S. (2013). Mathematical and pharmacokinetic modelling of epidermal and dermal transport processes. *Adv Drug Delivery Rev*, 65, 169-190.

Banyiova, K., Necasova, A., Kohoutek, J., Justan, I., & Cupr, P. (2016). New experimental data on the human dermal absorption of Simazine and Barbendazim help to refine the assessment of human exposure. *Chemosphere*, 145, 148-156.

Barnes N. (2010). Publish your computer code: It is good enough. *Nature*, 467, 753.

Becker, K., Conrad, A., Kirsch, N., Kolossa-Gehring, M., Schulz, C., Seiwert, M., & Seifert, B. (2007). German environmental survey (GerES): Human biomonitoring as a tool to identify exposure pathways. *Int J Hyg Environ Health*, *210*(3-4), 267-269.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res, 32*(Database issue), D267-270.

Boekelheide, K., & Campion, S. N. (2010). Toxicity testing in the 21st century: using the new toxicity testing paradigm to create a taxonomy of adverse effects. *Toxicol Sci, 114*(1), 20-24.

Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). PubChem: Integrated platform of small molecules and biological activities. *Annu Rep Comput Chem, 4*, 217-240.

Bornehag, C. G., & Nanberg, E. (2010). Phthalate exposure and asthma in children. *Int J Androl, 33*(2), 333-345.

Bornehag, C. G., Sundell, J., Weschler, C. J., Sigsgaard, T., Lundgren, B., Hasselgren, M., & Hagerhed-Engman, L. (2004). The association between asthma and allergic symptoms in children and phthalates in house dust: a nested case-control study. *Environ Health Perspect, 112*(14), 1393-1397.

Bracken, M. C., & Weiss, I. J. (1977). Database development in a regulatory agency. *J Chem Inf Comput Sci, 17*(4), 202-205.

Bridges, B. (2002). Fragrance: emerging health and environmental concerns. *Flavour and Fragrance Journal, 17*(5), 361-371.

Byer, W. L., Landau, H. B., Neufeld, M. L., & Rosenthal, H. (1976). Building a chemical ingredient data base for industrial and consumer products. *J Chem Inf Comput Sci, 16*(3), 137-141.

CDC. (2011). Fourth national report on human exposure to environmental chemicals. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Health Statistics.

Choi, H., Schmidbauer, N., Sundell, J., Hasselgren, M., Spengler, J., & Bornehag, C. G. (2010). Common household chemicals and the allergy risks in pre-school age children. *PLoS One, 5*(10), e13423.

Choudhury, H., Cogliano, J., Hertzberg, R., Mukerjee, D., Rice, G., Teuschler, L., . . . Schoeny, R. (2000). *Supplementary guidance for conducting health risk assessment for chemical mixtures*. (EPA/630/R-00/002). U.S. Environmental Protection Agency.

Chuprina, A., Lukin, O., Demoiseaux, R., Buzko, A., & Shivanyuk, A. (2010). Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J Chem Inf Model*, *50*(4), 470-479.

Cohen Hubal, E. A., Richard, A., Aylward, L., Edwards, S., Gallagher, J., Goldsmith, M. R., . . . Kavlock, R. (2010). Advancing exposure characterization for chemical evaluation and risk assessment. *J Toxicol Environ Health B Crit Rev, 13*(2-4), 299-313.

Colborn, T., vom Saal, F. S., & Soto, A. M. (1993). Developmental effects of endocrine-disrupting chemicals in wildlife and humans. *Environ Health Perspect, 101*(5), 378-384.

Comiskey, D., Api, A. M., Barratt, C., Daly, E. J., Ellis, G., McNamara, C., . . . Tozer, S. (2015). Novel database for exposure to fragrance ingredients in cosmetics and personal care products. *Regul Toxicol Pharmacol, 72*, 660-672.

Cowan-Ellsberry, C. E., & Robison, S. H. (2009). Refining aggregate exposure: example using parabens. *Regul Toxicol Pharmacol, 55*(3), 321-329.

Crisp, T. M., Clegg, E. D., Cooper, R. L., Wood, W. P., Anderson, D. G., Baetcke, K. P., . . . Patel, Y. M. (1998). Environmental endocrine disruption: An effects assessment and analysis. *Environ Health Perspect, 106*(Supplement 1), 11-56.

Csiszar, S. A., Ernstoff, A. S., Fantke, P., & Jolliet, O. (2017). Stochastic modeling of near-field exposure to parabens in personal care products. *J Expo Sci Environ Epidemiol*, 27, 152-159.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., . . . Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res, 36* (Database issue), D344-350.

Dellarco, V., Henry, T., Sayre, P., Seed, J., & Bradbury, S. (2010). Meeting the common needs of a more effective and efficient testing and assessment paradigm for chemical risk management. *J Toxicol Environ Health B Crit Rev, 13*(2-4), 347-360.

Dennis, K. K., Auerbach, S. S., Balshaw, D. M., Cui, Y., Fallin, M. D., Smith, M. T., Spira, A., Sumner, S., & Miller, G. W. (2016). The importance of the biological impact of exposure to the concept of the exposome. *Environ Health Perspect, 124*(10), 1504-1510.

Dennis, K. K., Marder, E., Balshaw, D. M., Cui, Y., Lynes, M. A., Patti, G. J., Rappaport, S. M., Shaughnessy, D. T., Vrijheid, M., & Barr, D. B. (2017). Biomonitoring in the era of the exposome. *Environ Health Perspect, 125*(4), 502-510.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297-302.

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., & Kavlock, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci, 95*(1), 5-12.

Dodson, R. E., Nishioka, M., Standley, L. J., Perovich, L. J., Brody, J. G., & Rudel, R. A. (2012). Endocrine disruptors and asthma-associated chemicals in consumer products. *Environ Health Perspect, 120*(7), 935-943.

DTSC. (2016). Safer Consumer Products (SCP). from https://www.dtsc.ca.gov/SCP/CandidateChemicalsList.cfm

Egeghy, P. P., Sheldon, L. S., Isaacs, K. K., Ozkaynak, H., Goldsmith, M. R., Wambaugh, J. F., Judson, R. S., & Buckley, T. J. (2016). Computational exposure science: An emerging discipline to support 21st-century risk assessment. *Environ Health Perspect*, 124(6), 697-702.

Egeghy, P. P., Vallero, D. A., & Cohen Hubal, E. A. (2011). Exposure-based prioritization of chemicals for risk assessment. *Environ Sci Policy, 14*(8), 950-964.

Ejaredar, M., Nyanza, E. C., Ten, E. K., & Dewey, D. (2015). Phthalate exposure and childrens neurodevelopment: A systematic review. *Environ Res*, *142*, 51-60.

Elobeid, M. A., & Allison, D. B. (2008). Putative environmental-endocrine disruptors and obesity: A review. *Curr Opin Endocrinol Diabetes Obes, 15*(5), 403-408.

Engel, S. M., Miodovnik, A., Canfield, R. L., Zhu, C., Silva, M. J., Calafat, A. M., & Wolff, M. S. (2010). Prenatal phthalate exposure is associated with childhood behavior and executive functioning. *Environ Health Perspect, 118*(4), 565-571.

EPA. (1986). *Guidelines for the health risk assessment of chemical mixtures*. (EPA/630/R-98/002). U.S. Environmental Protection Agency.

EPA. (2008). Toxicology testing in the 21st Century (Tox21). from http://www2.epa.gov/chemical-research/toxicology-testing-21st-century-tox21

EPA. (2010). Triclosan facts. from https://archive.epa.gov/pesticides/reregistration/web/html/triclosan_fs.html

EPA. (2015a). Triclosan: Response to petition. from https://www.epa.gov/ingredients-used-pesticide-products/triclosan#petition

EPA. (2015b). ToxCast and Tox21 summary files from invitrodb_v2. Retrieved from https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data on April 19, 2018. Data released October 2015.

Erythropel, H. C., Maric, M., Nicell, J. A., Leask, R. L., & Yargeau, V. (2014). Leaching of the plasticizer di(2-ethylhexyl)phthalate (DEHP) from plastic containers and the question of human exposure. *Appl Microbiol Biotechnol, 98*(24), 9967-9981.

EU. (2017). The EU list of potential endocrine disruptors. from http://eng.mst.dk/chemicals/chemicals-in-products/endocrine-disruptors/the-eu-list-of-potential-endocrine-disruptors/

Factor-Litvak, P., Insel, B., Calafat, A. M., Liu, X., Perera, F., Rauh, V. A., & Whyatt, R. (2014). Persistent associations between maternal prenatal exposure to phthalates on child IQ at age 7 years. *PLoS One, 9*(12), e114003.

Faulon, J. L., Brown, W. M., & Martin, S. (2005). Reverse engineering chemical structures from molecular descriptors: how many solutions? *J Comput Aided Mol Des, 19*(9-10), 637-650.

FDA. (2016). FDA issues final rule on safety and effectiveness of antibacterial soaps: Rule removes triclosan and triclocarban from over-the-counter antibacterial hand and body washes. from https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm517478.htm

FDA. (2017). Cosmetic labeling guide. from https://www.fda.gov/cosmetics/labeling/regulations/ucm126444.htm

Fonger, G. C., Hakkinen, P., Jordan, S., & Publicker, S. (2014). The National Library of Medicine's (NLM) Hazardous Substances Data Bank (HSDB): Background, recent enhancements and future plans. *Toxicology*, *325*, 209-216.

Fonger, G. C., Stroup, D., Thomas, P. L., & Wexler, P. (2000). TOXNET: A computerized collection of toxicological and environmental health information. *Toxicol Ind Health, 16*(1), 4-6.

Frasch, H. F. & Barbero, A. M. (2013). Application of numerical methods for diffusion-based modeling of skin permeation. *Adv Drug Delivery Rev*, 65, 208-220.

Frederiksen, M., Vorkamp, K., Jensen, N. M., Sorensen, J. A., Knudsen, L. E., Sorensen, L. S., Webster, T. F., & Nielsen, J. B. (2016). Dermal uptake and percutaneous penetration of ten flame retardants in a human skin ex vivo model. *Chemosphere*, 162, 308-314.

Frege, G. (1892). On sense and reference (M. Black, Trans.). *Translations from the philosophical writings of Gottlob Frege (P. Geach and M. Black, Eds.)*, pp 56-78, Basil Blackwell Oxford, 1960.

Gabb, H. A. & Blake, C. (2016a). An informatics approach to evaluating combined chemical exposures from consumer products: A case study for asthma-associated chemicals and potential endocrine disruptors. *Environ Health Perspect*, 124(8), 1155-1165.

Gabb, H. A. and Blake, C. (2016b). Response to "Comment on 'An informatics approach to evaluating combined chemical exposures from consumer products: A case study of asthma-associated chemicals and potential endocrine disruptors.'" *Environ Health Perspect*, 124, A156.

GAO. (2007). Comparison of U.S. and recently enacted European Union approaches to protect against the risks of toxic chemicals.

Geiger, R. S. & Ribes, D. (2010). The work of sustaining order in Wikipedia: The banning of a vandal. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*.

Glegg, G. A., & Richards, J. P. (2007). Chemicals in household products: problems with solutions. *Environ Manage, 40*(6), 889-901.

Goldsmith, M. R., Grulke, C. M., Brooks, R. D., Transue, T. R., Tan, Y. M., Frame, A., . . . Dary, C. C. (2014). Development of a consumer product ingredient database for chemical exposure screening and prioritization. *Food Chem Toxicol, 65*, 269-279.

Goldsmith, M. R., Tan, C., Chang, D. T., Grulke, C. M., Tornero-Velez, R., Vallero, D., . . . Phillips, L. (2012). Summary Report for "Personal chemical exposure informatics: Visualization and exploratory research in simulations and systems (PerCEIVERS)." EPA report EPA/600/R13/041.

Goodman, M., Lakind, J. S., & Mattison, D. R. (2014). Do phthalates act as obesogens in humans? A systematic review of the epidemiological literature. *Crit Rev Toxicol*, *44*, 151-175.

Grego, T., Pesquita, C., Bastos, H. P., & Couto, F. M. (2012). Chemical entity recognition and resolution to ChEBI. *ISRN Bioinformatics*.

Grun, F., & Blumberg, B. (2006). Environmental obesogens: Organotins and endocrine disruption via nuclear receptor signaling. *Endocrinology*, *147*(6 suppl), S50-S55.

Grun, F., & Blumberg, B. (2009). Endocrine disrupters as obesogens. *Mol Cell Endocrinol, 304*(1-2), 19-29.

Guo, Y., & Kannan, K. (2013). A survey of phthalates and parabens in personal care products from the United States and its implications for human exposure. *Environ Sci Technol, 47*(24), 14442-14449.

Hall, B., Steiling, W., Safford, B., Coroama, M., Tozer, S., Firmani, C., . . . Gibney, M. (2011). European consumer exposure to cosmetic products, a framework for conducting population exposure assessments Part 2. *Food Chem Toxicol, 49*(2), 408-422.

Hall, B., Tozer, S., Safford, B., Coroama, M., Steiling, W., Leneveu-Duchemin, M. C., . . . Gibney, M. (2007). European consumer exposure to cosmetic products, a framework for conducting population exposure assessments. *Food Chem Toxicol, 45*(11), 2097-2108.

Hand, E. (2010). People power. *Nature*, 466, 685-687.

Hansen, S., Lehr, C.-M., & Schaefer, U. F. (2013). Improved input parameters for diffusion models of skin absorption. *Adv Drug Delivery Rev*, 65, 251-264.

Harley K. G., Kogut, K., Madrigal, D. S., Cardenas, M., Vera, I. A., Meza-Alfaro, G., She, J., Gavin, Q., Zahedi, R., Bradman, A., Eskenazi, B., & Parra, K. L. (2016). Reducing phthalate, paraben, and phenol exposure from personal care products in adolescent girls: Findings from the HERMOSA intervention study. *Environ Health Perspect, 124*(10), 1600-1607.

Harris, Z. & Mattick Jr., P. (1988). Science sublanguges and the prospects for a global language of science. *Ann Am Acad Political Soc Sci*, 495, 73-83.

Hawizy, L., Jessop, D. M., Adams, N., & Murray-Rust, P. (2011). ChemicalTagger: A tool for semantic text-mining in chemistry. *J Cheminformatics, 3*(17).

Heindel, J. J. (2003). Endocrine disruptors and the obesity epidemic. *Toxicol Sci, 76*(2), 247-249.

Hengstler, J. G., Foth, H., Gebel, T., Kramer, P. J., Lilienblum, W., Schweinfurth, H., . . . Gundert-Remy, U. (2011). Critical evaluation of key evidence on the human health hazards of exposure to bisphenol A. *Crit Rev Toxicol, 41*(4), 263-291.

Hertz-Picciotto, I. & Delwiche, L. (2009). The rise of autism and the role of age at diagnosis. *Epidemiology*, *20*(1), 84-90.

Hettne, K. M., Stierum, R. H., Schuemie, M. J., Hendriksen, P. J., Schijvenaars, B. J., Mulligen, E. M., . . . Kors, J. A. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics, 25*(22), 2983-2991.

Hettne, K. M., van Mulligen, E. M., Schuemie, M. J., Schijvenaars, B. J., & Kors, J. A. (2010). Rewriting and suppressing UMLS terms for improved biomedical term identification. *J Biomed Semantics, 1*(5).

Ho, S.-M., Johnson, A., Tarapore, P., Janakiram, V., Zhang, X., & Leung, Y. K. (2012). Environmental epigenetics and its implication on disease risk and health outcomes. *ILAR J*, *53*(3-4), 363-373.

Houck, K. A., Judson, R. S., Knudsen, T. B., Martin, M. T., Richard, A. M., Crofton, K. M., Simeonov, A., Paules, R. S., Bucher, J. R., & Thomas, R. S. (2017). Comment on "On the utility of ToxCast™ and ToxPi as methods for identifying new obesogens." *Environ Health Perspect*, *125*(1), A8-A11.

Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S. A., Attene-Ramos, M., . . . Simeonov, A. (2016). Modelling the Tox21 10K chemical profiles for *in vivo* toxicity prediction and mechanism characterization. *Nat Commun, 7*(10425).

Humphreys, B. L., & Lindberg, D. A. (1993). The UMLS project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc, 81*(2), 170-177.

Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: An informatics research collaboration. *J Am Med Inform Assoc, 5*(1), 1-11.

IOM. (2014). Identifying and reducing environmental health risks of chemicals in our society: Workshop summary. *Roundtable on Environmental Health Sciences, Research, and Medicine*, Institute of Medicine, National Academies Press. from https://www.ncbi.nlm.nih.gov/books/NBK268889/

IRIS. (2017). EPA Integrated Risk Information System. from https://cfpub.epa.gov/ncea/iris_drafts/simple_list.cfm

Isaacs, K. K., Goldsmith, M. R., Egeghy, P., Phillips, K., Brooks, R., Hong, T., & Wambaugh, J. F. (2016). Characterization and prediction of chemical functions and weight fractions in consumer products. *Toxicol Rep*, 3, 723-732.

Isaacs, K. K., Phillips, K. A., Biryol, D., Dionisio, K. L., & Price, P. S. (2018). Consumer product chemical weight fractions from ingredient lists. *J Expo Sci Environ Epidemol*, 28, 216-222.

Janesick, A., & Blumberg, B. (2011). Endocrine disrupting chemicals and the developmental programming of adipogenesis and obesity. *Birth Defects Res C Embryo Today*, *93*(1), 34-50.

Janesick, A. S., Dimastrogiovanni, G., Chamorro-Garcia, R., & Blumberg, B. (2017). Reply to "Comment on 'On the utility of ToxCast™ and ToxPi as methods for identifying new obesogens.'" *Environ Health Perspect*, *125*(1), A12-A14.

Janesick, A. S., Dimastrogiovanni, G., Vanek, L., Boulos, C., Chamorro-Garcia, R., Tang, W., & Blumberg, B. (2016). On the utility of ToxCast™ and ToxPi as methods for identifying new obesogens. *Environ Health Perspect*, *124*(8), 1214-1226.

Jayjock, M. A., Chaisson, C. F., Franklin, C. A., Arnold, S., & Price, P. S. (2009). 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*(5), 515-524.

Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., & Murray-Rust, P. (2011). OSCAR4: a flexible architecture for chemical text-mining. *J Cheminform, 3*(1), 41.

Judson, R., Houck, K., Martin, M., Richard, A. M., Knudsen, T. B., Shah, I., Little, S., Wambaugh, J., Setzer, R. W., Kothya, P., Phuong, J., Filer, D., Smith, D., Reif, D., Rotroff, D., Kleinstreuer, N., Sipes, N., Xia, M., Huang, R., Crofton, K., & Thomas, R. S. (2016). Analysis of the effects of cell stress and cytotoxicity on *in vitro* assay activity across a diverse chemical and assay space. *Toxicol Sci*, *152*(2), 323-339.

Judson, R. S., Martin, M. T., Egeghy, P., Gangwal, S., Reif, D. M., Kothiya, P., Wolf, M., Cathey, T., Transue, T., Smith, D., Vail, J., Frame, A., Mosher, S., Cohen Hubal, E. A., & Richard, A. M. (2012). Aggregating data for computational toxicology applications: The U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) System. *Int J Mol Sci, 13*(2), 1805-1831.

Judson, R., Richard, A., Dix, D. J., Houck, K., Martin, M., Kavlock, R., Dellarco, V., Henry, T., Holderman, T., Sayre, P., Tan, S., Carpenter, T., & Smith, E. (2009). The toxicity data landscape for environmental chemicals. *Environ Health Perspect*, *117*(5), 685-695.

Kapraun, D. F., Wambaugh, J. F., Ring, C. L., Tornero-Velez, R., & Setzer, R. W. (2017). A method for identifying prevalent chemical combinations in the U.S. population. *Environ Health Perspect*, *125*(8), e098017.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., . . . Bryant, S. H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Res, 44*(D1), D1202-1213.

Klinger, R., Kolarik, C., Fluck, J., Hofmann-Apitius, M., & Friedrich, C. M. (2008). Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics, 24*(13), i268-276.

Knudsen T. B., Martin M. T., Kavlock R. J., Judson R. S., Dix D. J., & Singh A. V. (2009). Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB. *Reprod Toxicol*, *28*(2) 209–219.

Koniecki, D., Wang, R., Moody, R. P., & Zhu, J. (2011). Phthalates in cosmetic and personal care products: concentrations and possible dermal exposure. *Environ Res, 111*(3), 329-236.

Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2013). Overview of the chemical compound and drug name recognition (CHEMDNER) task. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop, 2*, 2-33.

Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2015a). CHEMDNER: The drugs and chemical names extraction challenge. *J Cheminform, 7* (Suppl 1 Text mining for chemistry and the CHEMDNER track), S1.

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., . . . Valencia, A. (2015b). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J Cheminform, 7* (Suppl 1 Text mining for chemistry and the CHEMDNER track), S2.

Krewski, D., Westphal, M., Andersen, M. E., Paoli, G. M., Chiu, W. A., Al-Zoughool, M., . . . Cote, I. (2014). A framework for the next generation of risk science. *Environ Health Perspect, 122*(8), 796-805.

Kripke, S. (1980). *Naming and Necessity,* Harvard University Press.

Kumar, P., Caradonna-Graham, V. M., Gupta, S., Cai, X., Rao, P. N., & Thompson, J. (1995). Inhalation challenge effects of perfume scent strips in patients with asthma. *Ann Allergy Asthma Immunol, 75*(5), 429-433.

Landau, H. B., & Byer, W. L. (1976). Production of a hierarchical chemical thesaurus. *J Chem Inf Comput Sci, 16*(3), 141-146.

Leaman, R., Wei, C. H., & Lu, Z. (2015). tmChem: a high performance approach for chemical named entity recognition and normalization. *J Cheminform, 7* (Suppl 1 Text mining for chemistry and the CHEMDNER track), S3.

Leigh, J. (2012). Systematic and trivial nomenclature. *Chemistry International, 34*(5), 28.

Loretz, L. J., Api, A. M., Babcock, L., Barraj, L. M., Burdick, J., Cater, K. C., . . . Scrafford, C. G. (2008). Exposure data for cosmetic products: facial cleanser, hair conditioner, and eye shadow. *Food Chem Toxicol, 46*(5), 1516-1524.

Loretz, L., Api, A. M., Barraj, L., Burdick, J., Davis de, A., Dressler, W., . . . Vater, S. (2006). Exposure data for personal care products: hairspray, spray perfume, liquid foundation, shampoo, body wash, and solid antiperspirant. *Food Chem Toxicol, 44*(12), 2008-2018.

Loretz, L. J., Api, A. M., Barraj, L. M., Burdick, J., Dressler, W. E., Gettings, S. D., . . . Sewall, C. (2005). Exposure data for cosmetic products: lipstick, body lotion, and face cream. *Food Chem Toxicol, 43*(2), 279-291.

Lowe, D. M., Corbett, P. T., Murray-Rust, P., & Glen, R. C. (2011). Chemical name to structure: OPSIN, an open source solution. *J Chem Inf Model, 51*(3), 739-753.

Martin, M. T., Judson, R. S., Reif, D. M., Kavlock, R. J., & Dix, D. J. (2009a). Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. *Environ Health Perspect, 117*(3), 392-399.

Martin M. T., Kavlock R. J., Rotroff D., Corum D., Judson R. S., & Dix D. J. (2009b). Profiling the reproductive toxicity of chemicals from multigeneration studies in the Toxicity Reference Database (ToxRefDB). *Toxicol Sci*, *110*(1):181–190.

McCray, A. T., Bodenreider, O., Malley, J. D., & Browne, A. C. (2001). Evaluating UMLS strings for natural language processing. *Proc AMIA Symp*, 448-452.

McNamara, C., Rohan, D., Golden, D., Gibney, M., Hall, B., Tozer, S., . . . Steiling, W. (2007). Probabilistic modelling of European consumer exposure to cosmetic products. *Food Chem Toxicol, 45*(11), 2086-2096.

Meeker, J. D. (2012). Exposure to environmental endocrine disruptors and child development. *Arch Pediatr Adolesc Med*, *166*(6), 952-958.

Montes-Grajales, D., & Olivero-Verbel, J. (2015). EDCs DataBank: 3D-Structure database of endocrine disrupting chemicals. *Toxicology, 327*, 87-94.

Muir, D. C., & Howard, P. H. (2006). Are there other persistent organic pollutants? A challenge for environmental chemists. *Environ Sci Technol, 40*(23), 7157-7166.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Comput Surv, 33*(1), 31-88.

Newbold, R. R. (2010). Impact of environmental endocrine disrupting chemicals on the development of obesity. *Hormones, 9*(3), 206-217.

Newbold, R. R., Padilla-Banks, E., Jefferson, W. N., & Heindel, J. J. (2008). Effects of endocrine disruptors on obesity. *Int J Androl, 31*(2), 201-208.

Nicole, W. (2018). Advocates for children's health: Working together to reduce harmful environmental exposures. *Environ Health Perspect*, *126*(1), e012001.

North, M. L., Takaro, T. K., Diamond, M. L., & Ellis, A. K. (2014). Effects of phthalates on the development and expression of allergic disease and asthma. Ann Allergy Asthma Immunol, 112, 496-502.

NRC. (2007). *Toxicity testing in the 21st century: A vision and a strategy*. Washington, D.C.: The National Academies Press.

Ott, W. R. (1990). Total human exposure: Basic concepts, EPA field studies, and future research needs. *J Air Waste Manag Assoc*, *40*, 966-975.

Palanza, P., Morellini, F., Parmigiani, S., & vom Saal, F. S. (1999). Prenatal exposure to endocrine disrupting chemicals: Effects on behavioral development. *Neurosci Biobehav Rev*, *23*(7), 1011-1127.

Park, Y. H., Lee, K., Soltow, Q. A., Strobel, F. H., Brigham, K. L., Parker, R. E., Wilson, M. E., Sutliff, R. L., Mansfield, K. G., Wachtman, L. M., Ziegler, T. R., & Jones, D. P. (2012). High-performance metabolic profiling of plasma from seven mammalian species for

simultaneous environmental chemical surveillance and bioeffect monitoring. *Toxicology*, *295*(1-3), 47-55.

Perera, F. & Herbstman, J. (2011). Prenatal environmental exposures, epigenetics, and disease. *Reprod Toxicol*, *31*(3), 363-373.

Perrin, J. M., Bloom, S. R., & Gortmaker, S. L. (2007). The increase of childhood chronic conditions in the United States. *JAMA*, *297*(24), 2755-2759.

Pollack, A. Z., Mumford, S. L., Krall, J. R., Carmichael, A. E., Sjaarda, L. A., Perkins, N. J., Kannan, K., & Schisterman, E. F. (2018). Exposure to bisphenol A, chlorophenols, benzophenones, and parabens in relation to reproductive hormones in healthy women: A chemical mixture approach. *Environ Int*, 120, 137-144.

Pollock, T., Mantella, L., Reali, V., & deCatanzaro, D. (2017). Influence of tetrabromobisphenol A, with or without concurrent triclosan, upon Bisphenol A and estradiol concentrations in mice. *Environ Health Perspect*, *125*(8), e087014.

Potthast, M. (2010). Crowdsourcing a Wikipedia vandalism corpus. *Proceedings of the 2010 ACM Special Interest Group on Information Retrieval*.

Randic, M. (1984). On molecular identification numbers. *J Chem Inf Comput Sci, 24*, 164-175.

Rappaport, S. M., Barupal, D. K., Wishart, D., Vineis, D. P., Scalbert, A. (2014). The blood exposome and its role in discovering causes of disease. *Environ Health Perspect, 122*(8), 769-774.

Rice, D. & Barone Jr., S. (2000). Critical periods of vulnerability for the developing nervous system: Evidence from humans and animal models. *Environ Health Perspect*, *108*(suppl 3), 511-533.

Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M. T., Wambaugh, J. F., Knudsen, T. B., Kancherla, J., Mansouri, K., Patlewicz, G., Williams, A. J., Little, S. B., Crofton, K. M., & Thomas, R. S. (2016). ToxCast chemical landscape: Paving the road to 21st century toxicology. *Chem Res Toxicol*, *29*, 1225-1251.

Richard, A. M., Swirsky Gold, L., & Nicklaus, M. C. (2006). Chemical structure indexing of toxicity data on the Internet: Moving toward a flat world. *Curr Opin Drug Discov Devel, 9*(3), 314-325.

Richard, A. M., & Williams, C. R. (2002). Distributed structure-searchable toxicity (DSSTox) public database network: A proposal. *Mutation Research, 499*, 27-52.

Rogers, W. J., & Aronson, A. R. (2008). Filtering the UMLS metathesaurus for MetaMap. from http://skr.nlm.nih.gov/papers/references/filtering07.pdf

Rotroff, D. M., Dix, D. J., Houck, K. A., Knudsen, T. B., Martin, M. T., McLaurin, K. W., Reif, D. M., Crofton, K. M., Singh, A. V., Xia, M., Huang, R. & Judson, R. S. (2013). Using *in vitro* high throughput screening assays to identify potential endocrine-disrupting chemicals. *Environ Health Perspect*, *121*, 7-14.

Sanderson, H., Counts, J. L., Stanton, K. L., & Sedlak, R. I. (2006). Exposure and prioritization – human screening data and methods for high production volume chemicals in consumer products: amine oxides a case study. *Risk Anal, 26*(6), 1637-1657.

Sanderson, H., Greggs, W., Cowan-Ellsberry, C., DeLeo, P., & Sedlak, R. (2013). Collection and dissemination of exposure data throughout the chemical value chain: A case study from a global consumer product industry. *Human Ecol Risk Assess: Int J, 19*(4), 999-1013.

SCCS. (2015). *The SCCS notes of guidance for the testing of cosmetic ingredients and their safety evaluation (9th revision)*. (SCCS/1564/15). European Commission.

Schwartz, A. S., & Hearst, M. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput, 8*, 451-462.

Schmidt, C. W. (2016). TCSA 2.0: A new era in chemical risk management. *Environ Health Perspect*, *124*(10), 182-186.

Selevan, S. G., Kimmel, C. A., & Mendola, P. (2000). Identifying critical windows of exposure for children's health. *Environ Health Perspect*, *108*(suppl 3), 451-455.

Service, R. F. (2009). A new wave of chemical regulations ahead? *Science*, *325*, 692-693.

Sexton, K., & Hattis, D. (2007). Assessing cumulative health risks from exposure to environmental mixtures - three fundamental questions. *Environ Health Perspect, 115*(5), 825-832.

Sheldon, L. S., & Cohen Hubal, E. A. (2009). Exposure as part of a systems approach for assessing risk. *Environ Health Perspect, 117*(8), 1181-1194.

Sohn, S., Comeau, D. C., Kim, W., & Wilbur, W. J. (2008). Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics, 9*, 402.

Soto, A. M., & Sonnenschein, C. (2010). Environmental causes of cancer: Endocrine disruptors as carcinogens. *Nat Rev Endocrinol, 6*, 363-370.

Steinemann, A. (2015). Volatile emissions from common consumer products. *Air Quality Atmosphere and Health, 8*(3), 273-281.

Steinemann, A. (2016). Comment on "An informatics approach to evaluating combined chemical exposures from consumer products: A case study of asthma-associated chemicals and potential endocrine disruptors." *Environ Health Perspect*, 124, A155.

Steinemann, A. C., MacGregor, I. C., Gordon, S. M., Gallagher, L. G., Davis, A. L., Ribeiro, D. S., & Wallace, L. A. (2011). Fragranced consumer products: Chemicals emitted, ingredients unlisted. *Environmental Impact Assessment Review, 31*(3), 328-333.

Stodden V., Donoho, D., Formel S., et al. (2010). Reproducible research: Addressing the need for data and code sharing in computational science. *Comput Sci Eng*, September/October, 8-12.

Stodden V., Guo P., and Ma Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE*, 8, e67111.

Sugino, M., Hatanaka, T., Todo, H., Mashimo, Y., Suzuki, T., Kobayashi, M., Hosoya, O., Jinno, H., Juni, K., & Sugibayashi, K. (2017). Safety evaluation of dermal exposure to phthalates: Metabolism-dependent percutaneous absorption. *Toxicol Appl Pharmacol*, 328, 10-17.

Tate, F. A. (1967). Handling chemical compounds in information systems. *Annu Rev Inform Sci, 2*, 285-309.

TCSA. (1976). Toxic Substances Control Act. Public Law, S. 3149, 94-469. from https://www.govtrack.us/congress/bills/94/s3149

TEDX. (2017) The endocrine disruption exchange. from https://endocrinedisruption.org/interactive-tools/tedx-list-of-potential-endocrine-disruptors/search-the-tedx-list

Tice, R. R., Austin, C. P., Kavlock, R. J., & Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: A Tox21 update. *Environ Health Perspect, 121*(7), 756-765.

Tomasulo, P. (2002). ChemIDplus-super source for chemical and drug information. *Med Ref Serv Q, 21*(1), 53-59.

Wallace, L. A. (1991). Comparison of risks from outdoor and indoor exposure to toxic chemicals. *Environ Health Perspect*, *95*, 7-13.

Wambaugh, J. F., Setzer, R. W., Reif, D. M., Gangwal, S., Mitchell-Blackwood, J., Arnot, J. A., . . . Cohen Hubal, E. A. (2013). High-throughput models for exposure-based chemical prioritization in the ExpoCast project. *Environ Sci Technol, 47*(15), 8479-8488.

Wambaugh, J. F., Wang, A., Dionisio, K. L., Frame, A., Egeghy, P., Judson, R., & Setzer, R. W. (2014). High throughput heuristics for prioritizing human exposure to environmental chemicals. *Environ Sci Technol, 48*(21), 12760-12767.

Wang, N. C., Jay Zhao, Q., Wesselkamper, S. C., Lambert, J. C., Petersen, D., & Hess-Wilson, J. K. (2012). Application of computational toxicological approaches in human health risk assessment. I. A tiered surrogate approach. *Regul Toxicol Pharmacol, 63*(1), 10-19.

Wang, N. C., Venkatapathy, R., Bruce, R. M., & Moudgal, C. (2011). Development of quantitative structure-activity relationship (QSAR) models to predict the carcinogenic potency of chemicals. II. Using oral slope factor as a measure of carcinogenic potency. *Regul Toxicol Pharmacol, 59*(2), 215-226.

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems, 28*(4), 1-38.

Welshons, W. V., Nagel, S. C., & vom Saal, F. S. (2006). Large effects from small exposures. III. Endocrine mechanisms mediating effects of bisphenol A at levels of human exposure. *Endocrinology*, *147*(6 suppl), S56-S69.

Wild, C. P. (2005). Complementing the genome with an “exposome”: The outstanding challenge of environental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*, *14*, 1847-1850.

Wild, C. P. (2012). The exposome: From concept to utility. *Int J Epidemiol*, *41*, 24-32.

WHO/UNEP. (2013). *State of the science of endocrine disrupting chemicals - 2012* (A. Bergman, J. J. Heindel, S. Jobling, K. A. Kidd & R. T. Zoeller Eds.): United Nations Environment Programme and the World Health Organization.

Yang, C. Z., Yaniger, S. I., Jordan, V. C., Klein, D. J., & Bittner, G. D. (2011). Most plastic products release estrogenic chemicals: a potential health problem that can be solved. *Environ Health Perspect, 119*(7), 989-996.

Zoeller, R. T., Brown, T. R., Doan, L. L., Gore, A. C., Skakkebaek, N. E., Soto, A. M., Woodruff, T. J., & Vom Saal, F. S. (2012). Endocrine-disrupting chemicals and public health protection: A statement of principles from The Endocrine Society. *Endocrinology, 153*(9), 4097-4110.

Appendix: Supplemental Material

The results of this dissertation research are provided in three spreadsheets. "Supplemental Material.xlsx" contains the results of various validation steps, preliminary analyses, and the final prioritizations. This is the primary results file referenced throughout this dissertation. "Supplemental Material Individual Authoritative Lists.xlsx" contains the final results for the individual lists of target chemicals, as described in Chapter 6.3. "Supplemental Material CPDB ToxCast Assay Summary.xlsx" contains the preliminary analysis of ToxCast data, as described in Chapter 8.2.1. Because reproducibility is so critical to scientific research, preliminary data and processing software are provided in addition these results spreadsheets. The "Code and Data.zip" archive contains the files described in Table 2, Table 5, Table 6, Table 9, Table 11, Table 16, and Table 19. The raw data for this research consists of the consumer product webpages scraped from Drugstore.com. Unfortunately, as discussed in Chapter 1, disseminating these HTML files would violate their copyright. However, web scraping instructions are provided to generate similar datasets. The consumer product usage data described in Chapter 3.4 is also subject to terms of use that prevent dissemination. However, a similar dataset can be purchased from Kantar Worldpanel.