Graphical Abstract

Predicting emerging chemical content in consumer products using machine learning

Luka Lila Thornton, David E. Carlson, Mark R. Wiesner

Highlights

Predicting emerging chemical content in consumer products using machine learning

Luka Lila Thornton, David E. Carlson, Mark R. Wiesner

- A proof-of-concept machine learning model estimates ranges of chemical concentrations in consumer products.
- The quantitative structure activity relationship (QSAR)-like model can be applied to data-poor chemicals due to relatively easy-to-obtain inputs.
- Case study shows promise for applying machine learning to engineered nanomaterial data despite a small training dataset.
- Chemical functional use data and product data may provide additional insight for predicting exposure-related endpoints.
- Presents two different, customizable machine learning frameworks for training and evaluating chemical endpoint prediction models under data-poor and data-rich scenarios.

Predicting emerging chemical content in consumer products using machine learning

Luka Lila Thornton^{a,b,*}, David E. Carlson^{a,c}, Mark R. Wiesner^{a,b}

^a*Duke University, Department of Civil and Environmental Engineering, 121 Hudson Hall, Durham, NC 27708, USA*

^b*Center for the Environmental Implications of NanoTechnology (CEINT), USA*

^c*Duke University, Department of Biostatistics and Bioinformatics, Duke University Medical Center, 2424 Erwin Road, Suite 1102 Hock Plaza, Durham, NC 27710, USA*

Abstract

Chemical ingredients in consumer products are continually changing. To understand our exposure to chemicals and their consequent risk, we need to know their concentrations in products, or chemical weight fractions. Unfortunately, manufacturers rarely report comprehensive weight fraction data on product labels. The goal of this study was to evaluate the utility of machine learning strategies for predicting weight fractions when chemical constituent data are limited. A “data-poor” framework was developed and tested using a small dataset on consumer products containing engineered nanomaterials to represent emerging substances. A second, more traditional framework was applied to a “data-rich” product dataset comprised of bulk-scale organic chemicals for comparison purposes. Feature variables included chemical properties, functional use categories (e.g., antimicrobial), product categories (e.g., makeup), product matrix categories, and whether weight fractions were manufacturer-reported or experimentally obtained. Classification into three weight fraction bins was done using a random forest or nonlinear support vector classifier. An ablation study revealed that functional use data improved predictive performance when included alongside chemical property data, suggesting the utility of functional use categories in evaluating the safety and sustainability of emerging chemicals. Models could roughly stratify material-product observations into order of magnitude weight fractions with moderate success; the best of these achieved an average balanced accuracy of 73% on the nanomaterials product data. Framework comparisons also revealed a positive trend in sample size versus average balanced accuracy, suggesting great promise for machine learning approaches with continued

Abbreviations: CV, cross validation; EPA, United States Environmental Protection Agency; OECD, Organisation for Economic Co-operation and Development; PCA, principal component analysis; RBF, radial basis function; RFC, random forest classifier; SVC, support vector classifier; WF, weight fraction

*Corresponding author

Email address: thorn.luka@gmail.com (Luka Lila Thornton)

investment in chemical data collection.

Keywords: Exposure modeling, Chemical function, Nanomaterials, Artificial intelligence, Cheminformatics, Environmental Exposure, Nanotechnology

1. Introduction

Over the past few decades, the role that chemical exposure plays in the onset of disease has become increasingly evident [1, 2]. Consequently, consumers have become more cautious of product ingredients. The ever-evolving landscape of products on the market requires more efficient and predictive means for evaluating consumer exposure. Laboratory analysis alone cannot keep pace, making predictive modeling the forefront solution. While there is extensive literature on toxicity modeling efforts to aid chemical policy decisions (e.g., quantitative structure activity relationships to estimate chemical toxicity endpoints), groups such as the National Research Council are emphasizing the need for complementary exposure models [3].

Figure 1 shows what kinds of data may be used to estimate exposure to substances in consumer products for a specific receptor (e.g., human) and time frame (e.g., daily). The data flow has been adapted from other models to address emerging chemicals, or materials, for which we have limited information [4, 5, 6]. The flow boils down to four components: the weight concentration of the chemical within the product, herein referred to as weight fraction (WF); the fraction released from the product matrix; the amount of product used over time; and the fraction that is biologically accessible to the receptor. In regard to determining these components for emerging chemicals, WF is perhaps the most approachable based on available data, and is the focus of our study. WF presents a significant data gap, as manufacturers are only required to disclose it on product labels in select contexts [7, 8].

Here we present modeling strategies to predict important intermediary values for estimating emerging chemical exposure, specifically, WF, through the use of machine learning. To date, machine learning methods such as random forest classification have been used to predict WF for organic chemicals as in Isaacs et al. [8]. Our objective was to build on these efforts using predictive variables, or *features*, that are more easily attainable for emerging or data-poor substances. A machine learning framework suitable for training on sparse data was developed, optimized and tested using engineered nanomaterial product data as a case study. Nanomaterials are widely regarded as materials with at least one dimension in the nanoscale (1–100 nanometers). Nanomaterials can be synthesized with a variety of properties and, thus, are used in a variety of applications. Their diversity and the presence of WF data in the published literature and public databases made them a suitable choice for investigation. For comparison, a more traditional framework was optimized and tested on the data-rich organic chemical product dataset from Isaacs et al. [8]. Both framework pipelines utilized one of two popular prediction algorithms—a random forest classifier (RFC) or a support vector classifier with a radial basis function kernel (SVC-RBF)—to predict a WF class of low (0.0–0.0001), medium (0.0001–0.01), or high (0.01–1.0). We also identified attainable, predictive features for estimating the quantities of emerging chemicals in products that may not have been

previously considered. Our proof of concept modeling approach makes a strong case for the application of machine learning methods and strategic feature selection to help understand the potential impact of emerging chemicals on consumers.

2. Materials and methods

In this section, we provide the rationale behind the datasets and features we selected to predict WF. Details on data curation may be found in Appendix B. We then walk through each step of our machine learning frameworks as summarized in Figure 3. The frameworks feature a generalized pipeline described in Table 2, where data preprocessing and classification algorithms are applied. For WF prediction, random forest classification and nonlinear support vector classification algorithms were tested using both frameworks. The predictive capacities of feature groups were also evaluated.

The code for this study was written in the Python programming language [9] with extensive use of the Scikit-learn library for machine learning [10] and Matplotlib for visualizing results [11]. The code can be launched in a web-accessible virtual environment via <https://mybinder.org/v2/gh/LukaThorn/ml-weight-fraction/master>. Alternatively, all data and code are available for download in Appendix A. Files for replicating the coding environment are also included to facilitate research reproducibility. By simply cloning our Git repository and running `poetry install` after the Python and Poetry [12] software have been installed, all code can be implemented in an isolated virtual environment with the same libraries and dependencies that were used for development.

2.1. Dataset selection

Two primary datasets were selected for the study. First, data on products containing nanomaterials were selected to represent a group of data-limited emerging substances. Nanomaterials product data (112 observations) were collated from the PEN Consumer Products Inventory [13], the Dutch National Institute for Public Health and the Environment [14], and the literature [15, 16]. Second, the bulk-scale, primarily organic chemicals product dataset used in Isaacs et al. [8] was selected for its relatively large size and wider coverage of consumer product space. These data, along with corresponding chemical property data, were provided by the US Environmental Protection Agency (EPA) [17, 8, 18]. Over 16,000 observations of chemicals in products remained after data were cleaned and filtered for the scope of the study. Such a sample size is more amenable to a traditional machine learning framework, enabling contextualization of results from the data-poor framework designed for the nanomaterials product dataset (see Section 2.3). Plots summarizing the compositions of both datasets are included in Figure 2 (chemical property features) and Appendix C (all other features).

2.2. Feature selection

To build practical prediction models suitable for emerging substances, features were chosen that were both readily available for training models and likely to be available for future products. It was also pragmatic to strive for a small number of features due to the

small number of nanomaterials product observations [30, chapters 15 and 18]. The selected features can be divided into the following groups: chemical properties, chemical functional uses, product categories (general product categories and product types), and product matrix information. These groups of features along with the dependent *target* variable, WF, are summarized in Table 1 with examples and data sources.

2.2.1. Chemical properties

In the context of this article, “chemical properties” refer to any properties that may be used to describe a chemical, regardless of whether they are physical, chemical (measured by changing the chemical composition of a substance) or both. Product manufacturers often select chemicals for use in products based on these properties. For example, waterproofing agents are hydrophobic and are associated with low solubility in water and high octanol-water coefficients, among other qualities [31].

Chemical property features were selected based on the following criteria: (1) they were representative of chemical composition, (2) they were applicable to both organic and inorganic chemicals at the nano- and bulk- scales, (3) they were relatively easy to measure or estimate [32], and (4) they were publicly available for the majority of substances. The final chemical properties selected were stoichiometric molecular weight (grams per mole of the molecular formula), melting point (degrees Celsius), boiling point (degrees Celsius), and density (grams per cubic centimeter). These features are often used in computational pharmacology and toxicology models, as are other features based off of molecular formulae such as the frequency of atoms in compounds, or higher order chemical descriptors containing structural or conformational information [33]. To estimate the molecular weight of nanomaterials, the molecular weight of the chemical formula for each nanomaterial’s bulk-scale equivalent was used. This approximation is not directly analogous or as physically meaningful compared to molecular weights for bulk-scale materials [34]. However, molecular weight may be pertinent in the case of dissolving nanomaterials. Moreover, it helps fulfill criterion 1 with only one feature slot, whereas other options like atom counts could require dozens of feature slots. For example, metals tend to have higher molecular weights, making inorganic compounds more distinguishable from organic ones that can behave quite differently. Melting point, boiling point and density also help satisfy this role. Additionally, while boiling point is not a very relevant property metal and metal oxide nanomaterial products, it was retained to keep the model generalizable to other emerging chemical data input. Other physicochemical property data, such as solubility data collected under consistent experimental conditions, were too sparsely available for the nanomaterials dataset to include. Criteria 2 and 4 disqualify some quintessential nanomaterial descriptors like size and functionalization. Most nanomaterial WF observations did not have these data. There are machine learning strategies that can account for missing data, although they are not necessarily reasonable to employ when sample size is small and the majority of observations are missing data for a candidate feature. In short, retaining observations was prioritized over the comprehensiveness of the feature set.

2.2.2. Functional use

The functional use of a chemical, sometimes simply referred to as function, can be described as the intended purpose of a chemical within a product (e.g., antimicrobial, catalyst). Functional use may be reported in manufacturer material safety data sheets or ingredient disclosures, as well as databases like the EPA’s Chemical and Product Categories database [29]. All nanomaterials product observations were individually assigned functional uses based on the terminology developed by the Organisation for Economic Co-operation and Development (OECD) [27]. For the organics product data, functional uses were remapped by their unique chemical identifier from the EPA’s database [28] to the OECD functional uses. This was primarily to help foster the adoption of the universal terminology proposed by the OECD. Appendix B includes more detail on this step. Functional uses were then merged with (assigned to) organics WF data by a unique chemical identifier, per Isaacs et al. [8].

2.2.3. Product categories

Weight fractions of different chemicals within the same product and functional use categories are hypothesized to be relatively similar. These feature data were not tested in the WF model by Isaacs et al. [8]. Product use category identifiers were provided by the EPA for all nanomaterials and organics product data according to their four-tiered categorization scheme based on product names and descriptions from product websites [29]. Features from the “general product category” tier (e.g., personal care) and “product type” tier (e.g., face wash) were retained. Product types are subsets of general product categories. Precautions to eliminate redundancy between these features are described in Section 2.3.

2.2.4. Product matrices

Information about the product matrix was also extracted from product use category identifier data; this includes whether it was a formulation versus an article (i.e., objects, such as a food container or clothing item), or if it was a spray versus non-spray.

2.2.5. WF measurement

Rogers et al. [16] found that experimental versus manufacturer-claimed nanomaterial concentrations could be quite variable in some products, so this information was retained as a binary feature for all product data.

2.3. Machine learning frameworks

Figure 3 summarizes the two different machine learning frameworks used for this study. The frameworks provide a broad overview of how data are processed and passed between algorithms, and the decision steps to train and filter multiple model options down to the best ones for the context. The framework on the left was designed for a data-poor scenario, which is often the case for emerging substances data. This framework was applied using the nanomaterials product data as a case study. A more traditional machine learning framework suitable for a much larger dataset is depicted on the right. This framework was applied to the bulk-scale organic chemicals product dataset for comparison purposes. While the expected data input is quite different between frameworks, the steps within the frameworks do not differ with the exception of step 4, model evaluation (highlighted).

2.3.1. Binning label data (optional)

As previously mentioned, target labels were converted from numeric weight fractions into weight fraction bins (low at 0.0–0.0001; medium at 0.0001–0.01; and high at 0.01–1.0), similar to Isaacs et al. [8]. These ranges were selected to strike a balance between interpretability and distributing both the nanomaterials and organic chemicals product data relatively evenly. The “high” bin had significantly fewer samples for both datasets despite the large range of weight fractions it spans (Appendix E.1). To account for class imbalance during model assessment, balanced accuracy was selected over a more typical scoring metric, such as accuracy. *Balanced accuracy* averages the accuracy of each of the true classes; so, the expected balanced accuracy for randomly guessing in a three-class classification problem is one-third.

The nature of the label data will also affect the choice of estimation algorithm used in step 4 of the pipeline (section 2.3.3). Binning the label data therefore required the use of a classifier for prediction.

2.3.2. Hyperparameter tuning and selection

The preprocessing and prediction algorithms discussed in the pipeline (2.3.3) have a number of hyperparameters that can be adjusted to tailor models to the training data (e.g., the number of trees in a random forest estimator). Combinations of sets of reasonable hyperparameters—collectively referred to as the parameter grid—are fit and tested in this step using a stratified five-fold cross validation (CV) pipeline. *Cross validation* is a standard machine learning strategy to tune parameters and address overfitting. It consists of randomly splitting data observations into k -groups, or folds, then running the desired algorithm or pipeline k -times, each time holding out a different data fold for validation while using the rest for training. Models are then assessed by the average performance score across k -iterations. So, in step 3 of both frameworks, the combination of parameters that resulted in the best average balanced accuracy score for each classifier was selected for use in step 4: model evaluation (2.4). Selecting parameter values based on predictive performance is considered a *supervised* learning approach.

2.3.3. The pipeline

The *pipeline* depicted in steps 2 and 4 of Figure 3 is detailed in Table 2. With the exception of the first (optional) step, the algorithms selected for this study exist as classes in the Scikit-learn library for simple, sequential implementation using the `Pipeline` class. The pipeline steps were first applied on the training data, then the validation data (framework step 2) or test data (framework step 4) using the appropriate Scikit-learn methods listed in Table 2. `Pipeline` also supports application of pipelines across multiple cross validation folds, as with framework step 2 (both frameworks) and framework step 4 (data-poor framework only).

The first three steps in the pipeline are preprocessing steps—transformations performed on data prior to prediction. Pipeline step 1, *data augmentation*, is considered optional. The process and results are described in Appendix D; there were not enough nanomaterial WF

observations to adequately assess whether data augmentation had a significant impact on performance.

For the *normalization* step of the pipeline, data were rescaled using min-max scaling. This only affected numerical data (i.e., chemical property data). The rest of the data were binary. Scaling helps to minimize bias that may be introduced if data are at different unit scales or ranges.

Next, a common *feature reduction* technique known as principal component analysis (PCA) was used to eliminate some redundancy within and between feature groups. PCA reduces the dimensionality of feature data while preserving as much variance from the original dataset as possible. This was particularly important for, e.g., highly correlated chemical properties like melting point and boiling point, and product category features and product type features, given their hierarchical relationship. A supervised approach was used to select the number of principal components to retain, `n_components`, during hyperparameterization. Rather than including all possible `n_components` in the parameter grid for framework step 2, this list was narrowed down to `n_components` with a cumulative explained variance ratio between 0.75 and 0.90 (see Suhr [35]).

There are many other feature reduction algorithms that could be considered for this pipeline step. Such algorithms also typically require selection of the desired number of features in advance. PCA was selected for its wide renown and ease of implementation in scikit-learn’s `Pipeline` class. Other algorithms such as hierarchical agglomerative clustering seemed to require far more custom code to make them pipeline-friendly (e.g., automating the process of narrowing down hyperparameters for framework step 2, or relabeling features post-processing).

Finally, we applied one of two different prediction algorithms for the *estimation* step: an RFC [30, chapter 15], as used in Isaacs et al. [8], or a nonlinear SVC. An RFC—or more broadly, random forest estimator—is a more advanced version of a decision tree algorithm. Decision trees are comprised of a series of decisions or criteria (branches) that split data into groups based on an optimized ruling or threshold for one feature per branch (up to a certain depth). Random forest estimators are considered an improvement over decision tree estimators, as they aggregate the decisions of numerous decision trees. Support vector machines use what are known as kernel functions to effectively map features to a higher dimensional feature space. A linear function is then applied in this newly mapped space, resulting in a nonlinear classifier (or regressor) in the original feature space. Smola and Schölkopf provide a good introduction to their mathematical basis [36]. For our SVC, an RBF kernel was selected for its affinity for non-parametric data, its popularity in domain adaptation strategies (e.g., data augmentation) and its ease of application using Scikit-learn’s `Pipeline`. Our study also included a linear SVC in earlier pipeline designs for the data-poor framework. The linear SVC performed poorly in the hyperparameter tuning step compared with the two more complex classifiers and was therefore dropped from the case studies to save on computational time.

2.4. Model evaluation and selection

Framework step 4, model evaluation, differs significantly between the data-poor and data-rich frameworks. The distinctions lie in the test data and the inclusion or exclusion of cross validation in the machine learning pipeline.

In a more traditional machine learning strategy such as the data-rich framework presented here, a random portion of all collated data will be set aside at the start of model development. This hold-out data, or blind test data, will only be used to test model performance after the model(s) has gone through framework steps 1-3 and no additional adjustments are anticipated. In our data-rich framework case study, one quarter (4110 observations) of the curated organics product dataset was withheld for this purpose. The algorithms from the pipeline were sequentially fit using the data and optimized parameters from framework steps 2 and 3. It should be noted that the data were not split, as cross validation was not employed in step 4 of the data-rich framework. The trained model was then applied to the blind test data following the methods in the test phase of 2. This process was repeated for each classifier and alternate datasets (see Section 2.5).

When there is a paucity of data for endpoint prediction, it may not be possible to set aside blind test data. In the data-poor framework, this issue is addressed by reusing all nanomaterials product data for both training and testing within the same 5-fold stratified cross validation pipeline as before. For model evaluation, the cross validation pipeline was also repeated one hundred times using different random states, which control the way the data are split into folds. Balanced accuracy scores were averaged across folds and random states, theoretically reducing potential bias in expected performance results from having relatively few samples within each fold.

In addition to overall balanced accuracy, model performance within each weight fraction bin was also evaluated. Figure 4 shows confusion matrices overlayed with a heat map to easily interpret scores, displaying the fractions of predicted weight fraction bins as they relate to the true bins in a three by three matrix after weighting for imbalances between bins. A hypothetical perfect model would result in an identity matrix colored dark blue along the principal diagonal. Permutation tests were also run to evaluate potential performance optimism from data reuse [37]. Models trained using the data-poor machine learning framework were tested using randomly permuted and unpermuted label data, then compared using the Wilcoxon signed rank test (Table Appendix E.4).

2.5. Impact of training data size and feature groups

To gain a better understanding of the effect of training sample size on performance (Figure 5), the data-rich machine learning framework was applied to organics product datasets consisting of $n_{train} = [100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000, 10^4]$. The same blind test data ($n_{blind} = 4110$) was used to score balanced accuracy for all models. In a similar manner, ablation studies were used to evaluate the predictive contributions of different feature groups or features for each classifier and product ingredient dataset combination using their corresponding frameworks (Table 3).

3. Results and discussion

When manufacturers incorporate new or emerging substances into consumer products, often there are few physicochemical property data, exposure data, or hazard data to evaluate their risk. For example, under the United States Toxic Substances Control Act, when manufacturers submit a minimum of ninety days notice to the EPA for production or importation of new chemical substances, only extant, non-confidential business information is required in the report [38]. This means that chemical risk assessors could have little data to decide if an emerging substance use-case constitutes unreasonable risk to public or environmental health. It is therefore essential that risk assessors have access to prediction tools that can still perform even when input data are limited, incomplete, or unreliable.

3.1. Feature selection

While a model is only as reliable as the quality of the training data, strategic model design—particularly, careful selection of model features—can help ameliorate the issues that regulators and risk assessors may encounter with data-poor substances. Table 3 shows the ablation study results, or how well the data-poor and data-rich frameworks performed on their respective case study datasets comprising different combinations of feature groups. The highest average balanced accuracy score for each classifier and dataset iteration are emphasized in boldface font. Because the nanomaterials product dataset only consisted of 112 observations, case study results from the more robust data-rich framework serve as a bellwether. I.e., trends from the data-poor framework tests were considered notable if they were also present in the data-rich framework results.

The best performing model for the data-poor framework trained and tested on nanomaterials product data was the SVC-RBF using (1) chemical property features, (2) functional use features and (3) the “Was WF measured?” feature, scoring 73.2% average balanced accuracy. Likewise, this combination of features resulted in the highest balanced accuracy for both classifiers on the organic chemicals product dataset—69.3% and 68.0% for the RFC and SVC-RBF, respectively. Excluding the “Was WF measured?” feature made no difference in the data-rich framework results, as all WF data were manufacturer-reported.

Chemical properties (i.e., chemical and physical) are the quintessential descriptors for predicting exposure and hazard endpoints [33, 7]. Expectantly, these features performed almost as well on their own as any of the feature group combinations based on the average balanced accuracy across all bins. The added value of the functional use and “Was WF measured?” features is more evident in Figure 4, where performance is stratified by WF bin. SVCs trained with functional use features scored well in the high WF category in comparison with SVCs trained using chemical properties alone; more high WF values were correctly predicted as high and fewer high WF values were incorrectly predicted as low. It is important to perform well in the high WF category due to the role that exposure plays in risk assessment. Risk is a function of exposure and hazard, and by extrapolation, WF (see Figure 1). So, not identifying a product ingredient that is present in high concentrations could be detrimental to consumers if that substance is confirmed to be toxic.

One should not assume that the performance of the product category and matrix property features (Table 3) reflects their utility in predictive modeling. Matrix properties were selected by what limited matrix data could be gleaned from the EPA’s curated product use category identifiers: formulations (versus non-formulations) and sprays (versus non-sprays). The product use category identifiers were not designed to capture matrix information the way it is detailed in product analyses such as Oomen et al. [14] (e.g., solid, cream, gel, aerosol, spray, liquid). Furthermore, these data were often unavailable for products. It is possible, with enough sample data, that matrix categories of similar specificity may be more predictive. The opposite issue may be the case for the product categories, which were hierarchical and so numerous that they had to be narrowed down before data could undergo automated feature reduction techniques (see Appendix B). Moreover, the datasets used in our study were not representative of the breadth and variability of products in commerce, but of what data were publicly available and amenable to curation. Using a subsection of data-rich products might improve the balanced accuracy contributions of product category data.

3.2. Impact of training dataset size

The impact of the size of the training dataset on model performance is shown in Figure 5. Application of the data-rich machine learning framework to varying amounts of organic chemicals product data resulted in a positive correlation between average balanced accuracy and the number of observations (log-scale). Based on this trend, we would expect an increase in data availability within the domain of the present feature data to continue to improve model performance before eventually plateauing. This is consistent with the literature from other fields (e.g., Huang et al. [39]). This potential for improved predictivity suggests that more data would enable the use of a regressor in Step 4 of the pipeline (Table 2) to predict numeric weight fractions instead of bins. Most notably, extrapolating this trend to the nanomaterial product data indicates great promise for predicting the prolificacy of emerging substances. These findings motivate the continued investment in curating and collating consumer product and exposure data.

3.3. Selecting a framework and model(s)

We presented two different machine learning frameworks (Figure 3) as case studies with a small and a large product ingredient dataset. The frameworks are customizable, with the options to mitigate the difficulties associated with smaller datasets such as binning data for order-of-magnitude classification instead of regression, or augmenting training data using data from a related “source” dataset. We anticipate more frequent data-limitations when evaluating exposure or hazard for emerging substances.

Between the nanomaterials and organic chemicals product datasets, most features had relatively comparable data (see Appendix C). The chemical properties for the respective datasets, however, formed two very distinct clusters (Figure 2). We felt this juxtaposition would constitute selection of the data-poor framework for predicting nanomaterial product WFs instead of attempting to apply models trained on organic chemicals product data. The differences in chemical property data also shaped the optional data augmentation strategy described in Appendix D.

When data are limited or present a large amount of uncertainty, users may wish to develop multiple models based on different framework options. A consensus approach can then be used to consolidate multiple predictions for new data.

4. Conclusions

Our machine learning framework for predicting weight fractions of emerging substances in consumer products is a proof-of-concept. It builds off of the work presented in Isaacs et al. [8] by expanding modeling options for data-limited scenarios and strategic selection of predictive, yet relatively obtainable features. Predictions are subject to several constraints detailed in Isaacs et al. [8] that will continue to limit the scope and applicability of exposure evaluation tools should data reporting remain at the status quo. The models presented here and by Isaacs and colleagues, however, can already estimate weight fractions within a few orders of magnitude with reasonable accuracy. Moreover, the positive trends in predictivity seen with increasing dataset size imply that there is significant potential for growth in predicting emerging substance prolificacy. Such predictions can be used for prioritizing more in-depth laboratory analysis of chemical exposure and risk. With further refinement, these kinds of machine learning models could be integrated as modules into exposure evaluation tools (e.g., SHEDS-HT [5]) or expanded with the data suggested in Figure 1 to obtain conservative Fermi estimates for individual or population level chemical exposure.

We also found that functional use data provided additional insight over chemical properties in the SVC-RBF models when it came to identifying substances used in higher concentrations. This and their overall high predictivity suggests that they may be a useful lens through which to evaluate the safety of nano- and bulk-scale chemicals. For example, risk assessments of emerging chemicals may include *read-across* approaches that involve referencing the behavior of other chemicals, based on their physicochemical properties, to estimate risk [34]. Functional use categories could provide a supplementary option for grouping chemicals for read-across purposes; Tickner et al. [40] has suggested such an application. In addition, Phillips et al. [28] demonstrated their potential utility in chemical alternatives assessment for, e.g., safer ingredient substitution in products. If functional use is to be considered for chemical evaluation and regulation, standardized terminology suitable for data analyses and more evidence of its predictive utility are required.

In all, the features and modeling strategies we selected for this case study present a viable step toward efficient prediction of the WF of emerging substances in consumer products that could be applied toward estimating other important exposure and hazard endpoints.

5. Acknowledgments

Special thanks to William Boyes, Kristin Isaacs, Kathie Dionisio and Daniel Vallero for their support and guidance, and to Bradley Beach for his help curating nanomaterials product data. This work received funding from the National Science Foundation (NSF) under grant DGE-2022040, and NSF and EPA under NSF Cooperative Agreements EF-0830093 and DBI-1266252 in support of CEINT. Any opinions, findings, conclusions or recommendations

expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or the EPA. This work has not been subjected to EPA review; no official endorsement should be inferred.

6. Author contributions

Luka L. Thornton: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing- Original draft preparation, Visualization. **David E. Carlson:** Methodology, Software, Writing- Reviewing and editing, Supervision. **Mark R. Wiesner:** Writing- Reviewing and editing, Funding acquisition, Supervision.

References

- [1] T. T. Schug, A. Janesick, B. Blumberg, J. J. Heindel, Endocrine disrupting chemicals and disease susceptibility, *The Journal of Steroid Biochemistry and Molecular Biology* 127 (2011) 204–215. URL: <http://www.sciencedirect.com/science/article/pii/S096007601100166X>. doi:<https://doi.org/10.1016/j.jsbmb.2011.08.007>.
- [2] S. M. Rappaport, Genetic Factors Are Not the Major Causes of Chronic Diseases, *PLoS One* 11 (2016) e0154387. URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4841510/>. doi:10.1371/journal.pone.0154387.
- [3] National Research Council, Exposure Science in the 21st Century, The National Academies Press, Washington, DC, 2012. URL: <https://www.nap.edu/catalog/13507/exposure-science-in-the-21st-century-a-vision-and-a>. doi:10.17226/13507.
- [4] M. Wiesner, J. Amos, CEINT’s NanoPHEAT Project : Integrating Exposure and Toxicity Data to Build Risk Forecasting Tools, 2020. URL: <https://ncihub.org/resources/2329>.
- [5] K. K. Isaacs, W. G. Glen, P. Egeghy, M.-R. Goldsmith, L. Smith, D. Vallero, R. Brooks, C. M. Grulke, H. Özkaynak, SHEDS-HT: An Integrated Probabilistic Exposure Model for Prioritizing Exposures to Chemicals with Near-Field and Dietary Sources, *Environmental Science & Technology* 48 (2014) 12750–12759. URL: <http://dx.doi.org/10.1021/es502513w>. doi:10.1021/es502513w.
- [6] Agency for Toxic Substances and Disease Registry, Public Health Assessment Guidance Manual, U.S. Department of Health and Human Services Public Health Service, Atlanta, Georgia USA, 2005. URL: https://www.atsdr.cdc.gov/hac/phamanual/pdfs/phagm_{_}final1-27-05.pdf.
- [7] P. P. Egeghy, R. Judson, S. Gangwal, S. Mosher, D. Smith, J. Vail, E. A. C. Hubal, The exposure data landscape for manufactured chemicals, *Science of The Total Environment* 414 (2012) 159–166. URL: <http://www.sciencedirect.com/science/article/pii/S0048969711012393>. doi:<https://doi.org/10.1016/j.scitotenv.2011.10.046>.
- [8] K. K. Isaacs, M.-R. Goldsmith, P. Egeghy, K. Phillips, R. Brooks, T. Hong, J. F. Wambaugh, Characterization and prediction of chemical functions and weight fractions in consumer products, *Toxicology Reports* 3 (2016) 723–732. URL: <http://www.sciencedirect.com/science/article/pii/S2214750016300671>. doi:<http://dx.doi.org/10.1016/j.toxrep.2016.08.011>.
- [9] Python Software Foundation, Python Language Reference, version 3.7, 2018.
- [10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, Technical Report, 2011. URL: <http://scikit-learn.sourceforge.net>.
- [11] J. D. Hunter, Matplotlib: A 2D graphics environment, *Computing in Science & Engineering* 9 (2007) 90–95. doi:10.1109/MCSE.2007.55.
- [12] S. Eustace, 2021, Poetry: Python packaging and dependency management made easy, URL: <https://python-poetry.org/>.

- [13] Project on Emerging Nanotechnologies, Consumer Products Inventory, 2013. URL: <http://www.nanotechproject.org/cpi>.
- [14] A. G. Oomen, M. Bennink, J. G. M. van Engelen, A. Sips, Nanomaterials in Consumer Products: Detection, Characterisation and Interpretation, Technical Report, Dutch National Institute for Public Health and the Environment [RIVM], Bilthoven, Netherlands, 2011. URL: <http://rivm.openrepository.com/rivm/handle/10029/257129>.
- [15] A. Philippe, J. Košík, A. Welle, J.-M. Guigner, O. Clemens, G. E. Schaumann, Extraction and characterization methods for titanium dioxide nanoparticles from commercialized sunscreens, *Environmental Science: Nano* 5 (2018) 191–202. URL: <http://dx.doi.org/10.1039/C7EN00677B>. doi:10.1039/C7EN00677B.
- [16] K. R. Rogers, J. Navratilova, A. Stefaniak, L. Bowers, A. K. Knepp, S. R. Al-Abed, P. Potter, A. Gitipour, I. Radwan, C. Nelson, K. D. Bradham, Characterization of engineered nanoparticles in commercially available spray disinfectant products advertised to contain colloidal silver, *Science of the Total Environment* 619-620 (2018) 1375–1384. URL: <https://www.sciencedirect.com/science/article/pii/S0048969717332539#s0105>. doi:10.1016/j.scitotenv.2017.11.195.
- [17] K. L. Dionisio, A. M. Frame, M. R. Goldsmith, J. F. Wambaugh, A. Liddell, T. Cathey, D. Smith, J. Vail, A. S. Ernstoff, P. Fantke, O. Jolliet, R. S. Judson, Exploring consumer exposure pathways and patterns of use for chemicals in the environment, *Toxicology Reports* 2 (2015) 228–237. URL: <https://actor.epa.gov/>. doi:10.1016/j.toxrep.2014.12.009.
- [18] A. J. Williams, C. M. Grulke, J. Edwards, A. D. McEachran, K. Mansouri, N. C. Baker, G. Patlewicz, I. Shah, J. F. Wambaugh, R. S. Judson, A. M. Richard, The CompTox Chemistry Dashboard: a community data resource for environmental chemistry, *Journal of Cheminformatics* 9 (2017) 61. URL: <https://doi.org/10.1186/s13321-017-0247-6>. doi:10.1186/s13321-017-0247-6.
- [19] Nature’s Greatest Secret, Personal Care, 2019. URL: <https://www.naturesgreatestsecret.co.uk/personal-care>.
- [20] Nanografi Nano Technology, Tungsten Disulfide (WS₂) Nanopowder/Nanoparticles, Purity: 99.99+%, Size: 35-75 nm, Accessed 2018-09-01. URL: <https://nanografi.com/nanoparticles/tungsten-disulfide-ws2-nanopowder-nanoparticles-purity-99-99-size-35-75-nm/>.
- [21] US Research Nanomaterials Inc, Nanopowders, Accessed 2018-09-01. URL: <https://www.us-nano.com/nanopowders>.
- [22] Sigma Aldrich, Tungsten(IV) sulfide, 2012. URL: <https://www.sigmaaldrich.com/catalog/product/aldrich/790583>.
- [23] J. Meija, B. Coplen Tyler, M. Berglund, A. Brand Willi, P. De Bièvre, M. Gröning, E. Holden Norman, J. Irrgeher, D. Loss Robert, T. Walczyk, T. Prohaska, Atomic weights of the elements 2013 (IUPAC Technical Report), 2016. URL: <https://www.degruyter.com/view/j/pac.2016.88.issue-3/pac-2015-0305/pac-2015-0305.xml>. doi:10.1515/pac-2015-0305.
- [24] SkySpring Nanomaterials Inc, Titanium Oxide Nanoparticles, 2016. URL: https://ssnano.com/inc/sdetail/titanium{}_oxide{}_nanoparticles/215.
- [25] American Elements, Nanomaterials, 2017. URL: <https://www.americanelements.com/nanomaterials-nanoparticles-nanotechnology.html>.
- [26] U.S. Environmental Protection Agency, Chemistry Dashboard, 2017. URL: <https://comptox.epa.gov/dashboard/>.
- [27] Organisation for Economic Co-operation and Development, Internationally Harmonised Functional, Product and Article Use Categories ENV/JM/MONO(2017)14, Technical Report, Paris, 2017. URL: [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2017\)14&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2017)14&doclanguage=en).
- [28] K. A. Phillips, J. F. Wambaugh, C. M. Grulke, K. L. Dionisio, K. K. Isaacs, High-throughput screening of chemicals as functional substitutes using structure-based classification models, *Green Chemistry* 19 (2017) 1063–1074. URL: <http://dx.doi.org/10.1039/C6GC02744J>. doi:10.1039/C6GC02744J.
- [29] K. L. Dionisio, K. Phillips, P. S. Price, C. M. Grulke, A. Williams, D. Biryol, T. Hong, K. K. Isaacs, The Chemical and Products Database, a resource for exposure-relevant data on chemicals in

- consumer products, *Scientific Data* 5 (2018) 180125. URL: <https://doi.org/10.1038/sdata.2018.125>. doi:10.1038/sdata.2018.125.
- [30] T. Hastie, R. Tibshirani, J. Friedman, in: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 ed., Springer, 2009. URL: <https://web.stanford.edu/~hastie/ElemStatLearn/>. doi:10.1007/978-0-387-84858-7.
 - [31] E. Voutsas, Chapter 11 - Estimation of the Volatilization of Organic Chemicals from Soil, in: *Thermodynamics, Solubility and Environmental Issues*, Elsevier, Amsterdam, 2007, pp. 205–227. URL: <https://www.sciencedirect.com/science/article/pii/B9780444527073500136>. doi:<https://doi.org/10.1016/B978-044452707-3/50013-6>.
 - [32] K. Mansouri, C. M. Grulke, R. S. Judson, A. J. Williams, OPERA models for predicting physicochemical properties and environmental fate endpoints, *Journal of Cheminformatics* 10 (2018) 10. URL: <https://doi.org/10.1186/s13321-018-0263-1>. doi:10.1186/s13321-018-0263-1.
 - [33] Y.-C. Lo, S. E. Rensi, W. Torng, R. B. Altman, Machine learning in chemoinformatics and drug discovery, *Drug Discovery Today* 23 (2018) 1538–1546. doi:10.1016/j.drudis.2018.05.010.
 - [34] T. Puzyn, G. Verheyen, S. Van Miert, B. Xing, S. Iqbal, Q. Zhao, V. Lobaskin, G. Basei, A. G. Papadiamantis, Y. Cohen, *Nanochemoinformatics and Statistical Modeling*, in: A. Haase, F. Klaessig (Eds.), *EU US Roadmap Nanoinformatics 2030*, 2018, pp. 37–55. doi:10.5281/zenodo.1486012.
 - [35] D. D. Suhr, Principal component analysis vs. exploratory factor analysis, *SUGI 30 proceedings* 203 (2005) 230. URL: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.444.2964{%&rep=rep1{%&}type=pdf>.
 - [36] A. J. Smola, B. Schölkopf, A tutorial on support vector regression, *Statistics and Computing* 14 (2004) 199–222. URL: <https://doi.org/10.1023/B:STCO.0000035301.49549.88>. doi:10.1023/b:stco.0000035301.49549.88.
 - [37] M. Ojala, G. C. Garriga, Permutation tests for studying classifier performance, *Journal of Machine Learning Research* 11 (2010) 1833–1863. URL: <https://dl.acm.org/doi/10.5555/1756006.1859913>. doi:10.1109/ICDM.2009.108.
 - [38] Federal Register, United States Code of Federal Regulations Title 40 §720 - Premanufacture Notification, U.S. Government Printing Office, 2018. URL: https://ecfr.io/Title-40/pt40.33.720{#}se40.33.720{_}145.
 - [39] C. Huang, L. S. Davis, J. R. G. Townshend, An assessment of support vector machines for land cover classification, *International Journal of Remote Sensing* 23 (2002) 725–749. URL: <https://doi.org/10.1080/01431160110040323>. doi:10.1080/01431160110040323.
 - [40] J. A. Tickner, J. N. Schifano, A. Blake, C. Rudisill, M. J. Mulvihill, Advancing Safer Alternatives Through Functional Substitution, *Environmental Science & Technology* 49 (2015) 742–749. URL: <https://doi.org/10.1021/es503328m>. doi:10.1021/es503328m.

Figure 1: The scope of the present study within a hypothetical data flow for estimating consumer or ecological exposure to emerging chemicals (EC) in consumer products. Chemical, product, and receptor data in the blue Venn diagram inform more complex extrinsic values at their intersections, such as chemical functional uses—or, further along the data flow, chemical weight fractions in products. Variables in the four gray arrows can be used to estimate exposure dose within a specific time frame prior to crossing biological barriers. The present study tests how well the target variable, weight fraction, can be predicted for emerging substances in consumer products using the indicated features and whether richer data sources like data on organic chemicals used in products (dashed oval) can aid in prediction.

Table 1: The following data were collated for both engineered nanomaterials and organic chemicals used in consumer products in order to model maximum weight fraction, the target variable for prediction. Data groups are broken down into specific features, or predictive variables, and their data type prior to feature agglomeration. † Some or all nanomaterials consumer product data were curated by the authors.

Data group	Features or Target Variable	Data type	Sources
Weight fraction	Maximum weight fraction	Numerical converted to ordinal;	[14, 13, 8, 15, 16, 19]
	Was weight fraction measured (or reported by manufacturer)?	Binary	†
Chemical properties	Molecular weight; Melting point; Boiling point; Density	Numerical	[20, 21, 22, 23, 24, 25, 26]
Chemical functional uses	Absorbent; Antiadhesive / anticohesive; Antioxidant; Antiredeposition agent; Biocide; Brightener; Chemical reaction regulator; Colorant; Deodorizer; Filler; Flavouring / nutrient; Hardener; Humectant; Opacifier; UV stabilizer; Viscosity modifier	Binary	[27, 28] †
General product categories	Arts and crafts / office supplies; Cleaning products and household care; Electronics / small appliances; Home maintenance; Personal care; Pesticides; Vehicle	Binary	[29] †
Product types	Child specific; Dental care; Deodorant; Engine maintenance; Fabric treatment and dye; Facial cleansing and moisturizing; Fungicide; General household cleaning; General moisturizing; Makeup and related; Paint, stain and related products; Sunscreen; Surface sealers	Binary	[29] †
Product matrix	Formulation (or non-formulation); Spray (or non-spray)	Binary	[29] †

Figure 2: A three-dimensional scatterplot of physical chemical properties for inorganic nanomaterials and bulk-scale organic chemicals found in the consumer product datasets.

Table 2: Steps in the machine learning pipeline and how they are applied to training and validation/test data using methods in the Python-based Scikit-learn machine learning library. X represents a matrix of feature (descriptor) data; y represents a vector of label (predictive variable) data.

Pipeline step	Training phase		Validation / test phase	
	X	y	X	y
Data augmentation (optional)	fit, transform +mask	fit, transform		
Normalization	fit, transform		transform	
Feature reduction	fit, transform		transform	
Estimation	fit	fit	predict	predict

Data-poor Machine Learning Framework

Data-rich Machine Learning Framework

Figure 3: Overview of the machine learning approaches used to predict chemical weight fractions in consumer products for data-poor (left) and data-rich (right) scenarios. Data on products containing nanomaterials were used as a case study to represent an emerging chemicals product dataset with limited samples. Another dataset comprised of over 16,000 bulk-scale organic chemical weight fraction observations was used for the data-rich modeling scenario. The steps in each modeling strategy were repeated for two different classifiers, a random forest and a nonlinear support vector machine. Pipeline steps are detailed in Table 2.

Table 3: A comparison of how well different combinations of feature groups predict weight fraction (WF) categories using two different machine learning frameworks: one for a data-poor nanomaterials product dataset and one for a data-rich bulk-scale organic chemicals product dataset. Values represent average balanced accuracies for either a random forest classifier (RFC) or a support vector classifier (SVC) with a nonlinear kernel.

Chemical properties	Feature group				Data-poor		Data-rich	
	Func- tional uses	Product categories	Matrix properties	WF measured?	RFC	SVC	RFC	SVC
X					0.683	0.682	0.689	0.572
	X				0.657	0.728	0.587	0.578
		X			0.607	0.583	0.426	0.426
			X		0.664	0.666	0.389	0.389
X	X				0.657	0.661	0.693	0.680
X	X			X	0.648	0.732	0.693	0.680
X	X		X	X	0.648	0.712	0.674	0.655
X	X	X	X		0.577	0.650	0.624	0.616
X	X	X	X	X	0.621	0.649	0.617	0.616

Figure 4: A table of confusion matrices showing how well select groups of features performed in data-rich and data-poor scenarios using a random forest classifier (RFC) and a nonlinear support vector classifier (SVC). The scores shown are average balanced accuracies across low, medium and high weight fraction bins. The values along the principal diagonals of matrices are averaged to get the overall balanced accuracies displayed in Table Appendix E.3.

Figure 5: Scatterplot of performance (average balanced accuracy) for two classifiers, a random forest classifier (RFC) and a nonlinear support vector classifier (SVC), versus sample size (log-scale). Classifiers were trained on bulk-scale organic chemicals product data to predict a chemical weight fraction category for blind test data.