Potential for Machine Learning to Address Data Gaps in Human Toxicity and Ecotoxicity Characterization

von Borries, Kerstin; Holmquist, Hanna; Kosnik, Marissa; Beckwith, Katie V.; Jolliet, Olivier; Goodman, Jonathan M.; Fantke, Peter

Published in:
Environmental Science and Technology

Link to article, DOI:
10.1021/acs.est.3c05300

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Environmental Science & Technology

pubs.acs.org/est

Article

Potential for Machine Learning to Address Data Gaps in Human Toxicity and Ecotoxicity Characterization

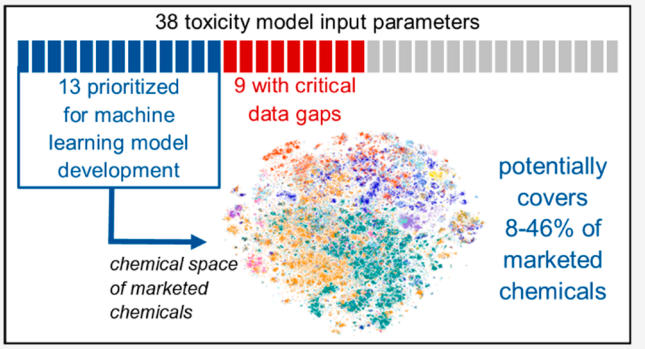
Kerstin von Borries,* Hanna Holmquist, Marissa Kosnik, Katie V. Beckwith, Olivier Jolliet, Jonathan M. Goodman, and Peter Fantke*

Cite This: *Environ. Sci. Technol.* 2023, 57, 18259−18270

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine Learning (ML) is increasingly applied to fill data gaps in assessments to quantify impacts associated with chemical emissions and chemicals in products. However, the systematic application of ML-based approaches to fill chemical data gaps is still limited, and their potential for addressing a wide range of chemicals is unknown. We prioritized chemical-related parameters for chemical toxicity characterization to inform ML model development based on two criteria: (1) each parameter's relevance to robustly characterize chemical toxicity described by the uncertainty in characterization results attributable to each parameter and (2) the potential for ML-based approaches to predict parameter values for a wide range of chemicals described by the availability of chemicals with measured parameter data. We prioritized 13 out of 38 parameters for developing ML-based approaches, while flagging another nine with critical data gaps. For all prioritized parameters, we performed a chemical space analysis to assess further the potential for ML-based approaches to predict data for diverse chemicals considering the structural diversity of available measured data, showing that ML-based approaches can potentially predict 8−46% of marketed chemicals based on 1−10% with available measured data. Our results can systematically inform future ML model development efforts to address data gaps in chemical toxicity characterization.



KEYWORDS: *prioritization, uncertainty, chemical space, chemical properties, life cycle impact assessment, chemical substitution, risk screening, safe and sustainable by design*

1. INTRODUCTION

Global chemical production and use are increasing with growing trends in urbanization, economic growth, and living standards.[1] To fully benefit from the positive contribution of chemicals to society, such as ensuring food security and healthcare, rigorous chemical assessment and management are crucial to preventing unintended chemical impacts on humans and ecosystems. However, assessing the growing number of marketed chemicals across different consumer products, populations, and environments is increasingly challenging.[2,3] Characterizing chemical toxicity impacts, including aspects on environmental fate, exposure, and (eco-)toxicity effects, is essential across a wide range of chemical-related decision support tools, such as risk assessment[4,5] and screening,[6] life cycle impact assessment (LCIA),[7,8] chemical footprinting,[9,10] chemical substitution,[11] benchmarking chemical pollution against local-to-global boundaries,[12,13] and safe-and-sustainable-by-design (SSbD) assessments.[14] The application of chemical-related decision support tools to the >100,000 marketed chemicals[15] and the wide range of product uses is currently limited by a lack of structured, high-quality input data needed to characterize toxicity for millions of chemical−product combinations.[16,17] Obtaining new data from experimental tests is cost- and time-consuming, and confidential or nontransparent reporting hinder access to existing data.[18−20] To address data gaps, scientists have been developing quantitative structure−activity relationships (QSAR) for decades by creating quantitative links between chemical structures and various target properties, including input parameters for characterizing chemical toxicity.[21,22] With increasing data availability and computing power, QSAR evolved from simple regressions on small sets of congeneric compounds to applying advanced statistical and machine learning (ML) techniques on large chemical sets with diverse molecular structures, boosting their predictive performance and applicability for a broader realm of chemicals.[23,24] Several

Special Issue: Data Science for Advancing Environmental Science, Engineering, and Technology

Received: July 14, 2023
Revised: October 12, 2023
Accepted: October 13, 2023
Published: November 1, 2023

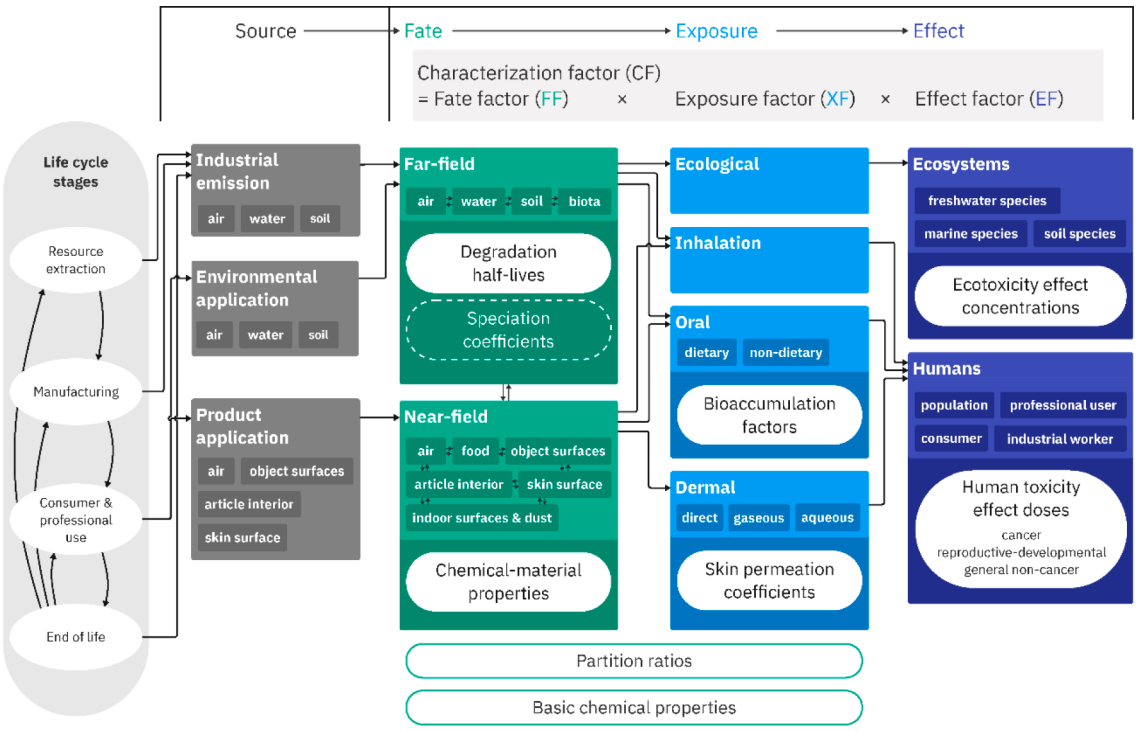


Figure 1. Schematic illustration of nine groups of chemical-related parameters (white pill-shaped boxes) needed for characterizing toxicity with USEtox from chemical emission sources covering process emissions, direct environmental application (e.g. pesticide spray applications), and application in products (dark gray column) via fate within and across near-field and far-field environmental compartments* (green column) and exposure to contaminated compartments following different exposure routes (light blue column) to harmful effects on humans and ecosystems** (dark blue column). *Small arrows indicate chemical mass exchange across compartments; not all possible exchange pathways are shown. Degradation describes biotic and abiotic degradation processes; in-biota metabolism and the formation of metabolites are currently not directly considered within USEtox. **USEtox models effects on humans and other species based on intake dose-level effect data that include internal metabolic processes.

advanced chemical data prediction models are readily accessible through public modeling suites providing predictions for multiple chemical properties[25,26] and many more have been documented in the scientific literature for individual chemical properties, including dissociation constants,[27,28] root concentration factors,[29−31] and ecotoxicity end points.[32−37] While the development of ML-based approaches has been an active field of research, a systematic adaptation for chemical toxicity characterization is still limited. The main challenges relate to a lack of oversight into required input parameters which could support the systematic development of ML-based approaches and a lack of transparency about whether such approaches can robustly predict parameter values for diverse chemicals.

To address these limitations, the main goal of this study was to prioritize chemical toxicity characterization parameters for developing ML-based approaches and assess the potential of these approaches to fill input data gaps for a wide range of marketed chemicals. To achieve this goal, we defined three specific objectives: (a) to propose and apply a framework considering characterization results uncertainty and data availability for prioritizing chemical-related input parameters in toxicity characterization frameworks for which ML-based approaches are most relevant, (b) to explore trends in characterization uncertainty across chemicals and source compartments for assessing parameter relevance in different scenarios, and (c) to analyze the chemical space covered by available data for assessing the potential of ML-based approaches to predict prioritized parameters for diverse chemical structures.

2. METHODS

2.1. Characterizing Chemical Toxicity. Modeling frameworks that characterize human toxicity and ecotoxicity impacts are built on various input parameters. Many of these parameters are chemical- or chemical-product-specific. However, measured data or appropriate estimates are lacking for most marketed chemicals,[15] leading to substantial data gaps across toxicity characterization frameworks. To prioritize relevant parameters for filling data gaps with ML-based approaches, we assessed 38 chemical-related parameters used in the global scientific consensus modeling framework USEtox.[38] We focused on USEtox as a reference framework since it is widely applied in comparative toxicity impact assessment,[39−42] considers human toxicity and ecotoxicity impacts, and covers different exposure routes and emission- and product-based impact pathways.[7,8] USEtox expresses toxicity impacts through characterization factors (CFs) combining environmental fate, exposure, and effects based on mass balance principles. These factors depend on underlying parameters, such as partition ratios, environmental half-lives, and intake-related toxicity effect doses. Figure 1 summarizes all chemical-related parameters in nine groups and outlines their primary relevance in modeling fate, exposure, and effect with USEtox (see Supporting Information (SI), Section S1.1a for details). Speciation coefficients were outside the scope of our analysis due to their exclusive applicability for the small group

https://doi.org/10.1021/acs.est.3c05300
Environ. Sci. Technol. 2023, 57, 18259−18270

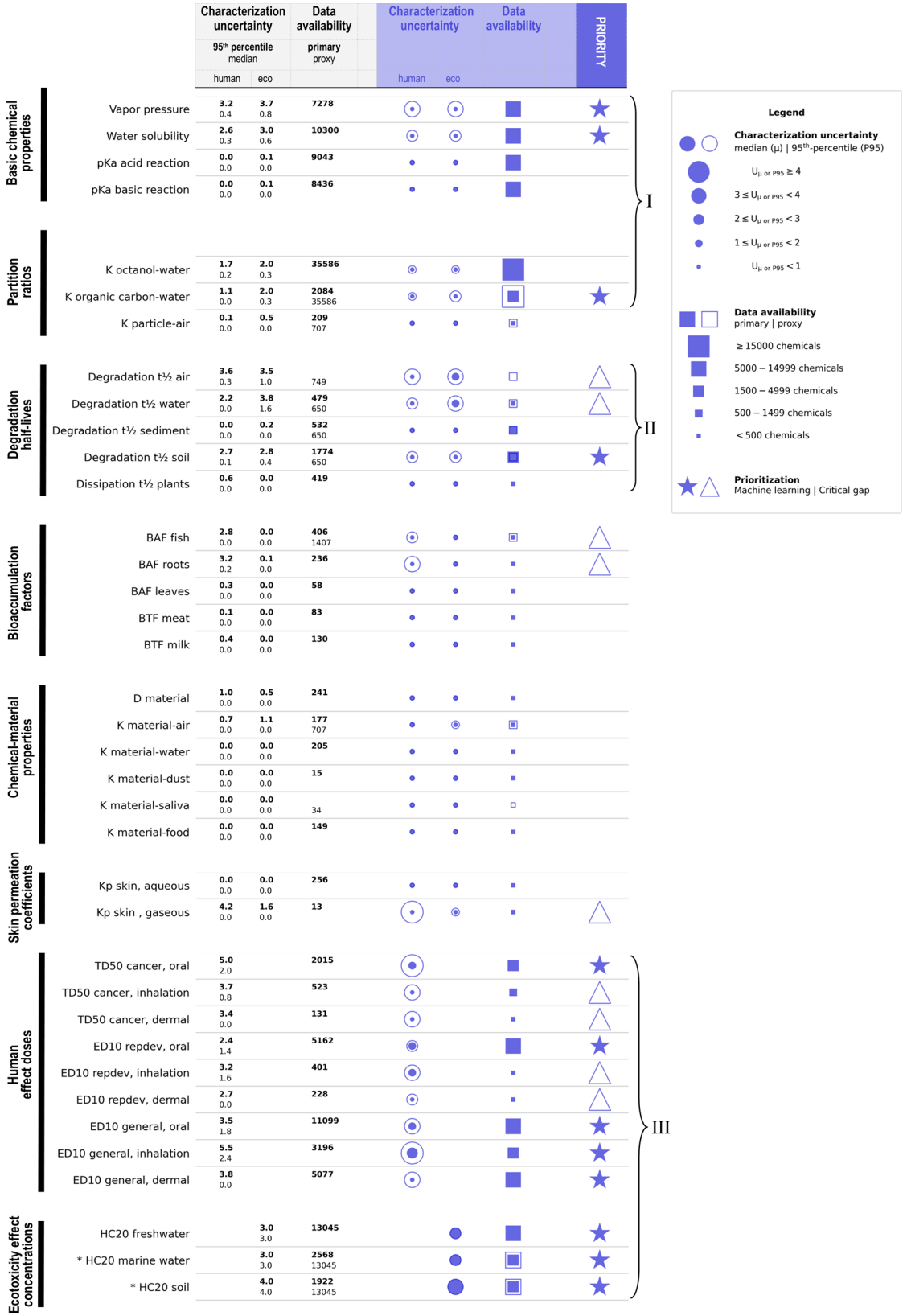


Figure 2. Assessment of 38 chemical-related toxicity characterization input parameters for characterization uncertainty (●) and data availability (■), used to assign prioritization flags for machine learning model development (★) and flag critical gaps (△) with three highlighted parameter groups (I–III). Characterization uncertainty and data availability are presented as nested symbols featuring the median (inner) and 95th percentile (outer) for the uncertainty and chemicals with measured data based on primary (inner) and proxy (outer) data sources for the data availability. The first two columns present the data underlying the characterization uncertainty and data availability classification. K = partition ratio, $t^1/_2$ = half-life, BAF = bioaccumulation factor, BTF = biotransfer factor, D = diffusion coefficient, Kp = skin permeation coefficient, TD50 = 50% response tumoric dose, ED10 = 10% response effect dose, HC20 = 20% response hazardous concentration, U = absolute characterization uncertainty distribution.

https://doi.org/10.1021/acs.est.3c05300
Environ. Sci. Technol. 2023, 57, 18259−18270

of metal ions and therefore do not require ML-based approaches applicable to a wide range of chemical structures.[43]

2.2. Framework for Prioritizing Parameters for Machine Learning. While various parameters are needed to characterize toxicity impacts, they are neither equally relevant for obtaining robust characterization results nor equally well-suited for developing ML-based prediction methods. We proposed a framework to prioritize parameters for ML model development based on two criteria: (1) *uncertainty* in toxicity characterization results attributable to each parameter and (2) *data availability* given by the number of chemicals with measured parameter data. Characterization uncertainty determines how strongly input parameter uncertainty affects the characterization results, which helps identify the most relevant parameters for obtaining robust results. Data availability is a limiting factor for developing ML-based approaches since such methods are principally more predictive when abundant, diverse training data are available.[44] Hence, data availability informs us where ML has the potential to make predictions for a wide range of chemicals. Parameters were prioritized if they exceeded a "medium" uncertainty and data availability class assigned on a five-point scale, corresponding to a minimum uncertainty of 2 orders of magnitude and data availability for at least 1500 chemicals (see SI, Section S1.2a).

2.2.1. Uncertainty Analysis. To determine the relevance of each parameter for obtaining robust characterization results, we performed an uncertainty analysis using an adapted version of USEtox 3.0 beta (https://usetox.org, see SI, section S1.3a). Each parameter was modified within its 95% confidence interval derived from parameter-specific squared geometric standard deviations, GSD^2, which we defined based on performance reported in the literature for available parameter prediction models commonly used or integrated with USEtox (see SI, Table S2-1). By propagating this input parameter uncertainty for 3421 chemicals and ten source compartments covering emissions to near- and far-field compartments, we obtained a distribution of characterization results uncertainty for each parameter, reflecting the nonlinearity in USEtox modeling behavior (see SI, Section S1.3b). To assign each parameter an uncertainty class on our five-point scale (see SI, Table S1-2), we derived the median and 95th percentile of the absolute uncertainty distribution across these chemical-emission scenarios. Parameter prioritization was based on the 95th percentile as it allowed capturing strong nonlinearity or interactions with other parameters, while the median served as a reference point describing central tendency. The underlying characterization uncertainty distribution was further analyzed to identify trends with the parameter value and source compartment that can feed into ML model development.

2.2.2. Data Availability and Chemical Space Analysis. To assess the quantity and chemical diversity of data available for training ML-based approaches, we identified chemicals for which measured parameter data were available in large public data repositories and curated data sets published in the scientific literature (see SI, section S1.4a and Table S2-2). To assign a data availability class on our 5-point-scale (see SI, Table S1-3), we derived the number of unique chemicals with measured data across identified data repositories using InChIKeys as chemical identifiers to harmonize chemical identification. For prioritized parameters, we performed a chemical space analysis to assess how well chemicals with measured data represent the structural diversity of the wider realm of chemicals using the existing space of marketed chemicals as a reference. We compiled this space of marketed chemicals from chemical lists provided through U.S. EPA's CompTox Chemicals Dashboard v2.1 (https://comptox.epa.gov/dashboard/),[45] focusing on official registration and specific applications lists (see SI, Section S1.4b). All marketed chemicals were categorized into chemical classes using the *ClassyFire* chemical taxonomy.[46] We visualized the space of marketed chemicals by mapping Morgan fingerprints[47] calculated with RDKit v2022.3.5[48] into two dimensions using a t-distributed Stochastic Neighbor Embedding[49,50] (t-SNE) (see SI, Section S1.4c). As an estimate of the predictive potential of ML-based approaches, we defined a structural domain for each prioritized parameter based on the similarity of every marketed chemical with its five nearest neighbors among chemicals with measured data derived from average Jaccard distances (see SI, Section S1.4d). We note that the presented structural domains provide no information on the accuracy a model may achieve for chemicals considered "inside" vs "outside" the domain, as the choice of ML algorithms and appropriate training features impacts the achievable performance.[51]

Based on our analysis, we discussed the potential and limits of ML-based approaches for addressing the prioritized parameters and systematically closing gaps in chemical toxicity characterization. All analysis and visualization steps were done with Python 3.10[52] (see SI, Sections S1.3f and S1.4f for details).

3. RESULTS AND DISCUSSION

3.1. Prioritized Parameters for Machine Learning. We assessed the priority of 38 input parameters from the USEtox toxicity characterization framework for developing ML-based prediction methods based on the uncertainty propagated into human toxicity and ecotoxicity characterization results and the data availability for each input parameter (Figure 2). In brief, the higher the uncertainty class of a parameter, the more important it is for obtaining robust characterization results, and the higher its data availability class, the higher its potential to train ML-based approaches for predicting a wide range of chemicals. Parameters with "moderate", "high", or "very high" uncertainty and data availability classes were flagged for prioritization of ML development. Parameters that show high relevance for robust toxicity characterization (indicated by the uncertainty) but insufficient data available for developing ML-based approaches received a triangular flag symbol, indicating a critical data gap.

Based on our analysis, 13 parameters were prioritized for ML model development, while another nine were flagged as critical gaps due to low data availability. Among the 13 prioritized parameters, seven parameters showed high or very high uncertainty (3.2 to 5.5 orders of magnitude, OOM), and six parameters showed high data availability (5077 to 13045 chemicals with measured data). In the following, we summarize and discuss three groups of flagged parameters related to environmental partitioning, environmental degradation, and toxicity effects.

3.1.1. Environmental Partitioning. The first parameter group relates to environmental partitioning, namely, Henry's law constant with its underlying water solubility and vapor pressure, and the organic carbon−water partition ratio (K_{OC}). Characterization uncertainty attributable to these parameters was higher for ecotoxicity than for human toxicity impacts, reaching high (3.7 OOM) and very high (4.8 OOM)

https://doi.org/10.1021/acs.est.3c05300
Environ. Sci. Technol. 2023, 57, 18259−18270

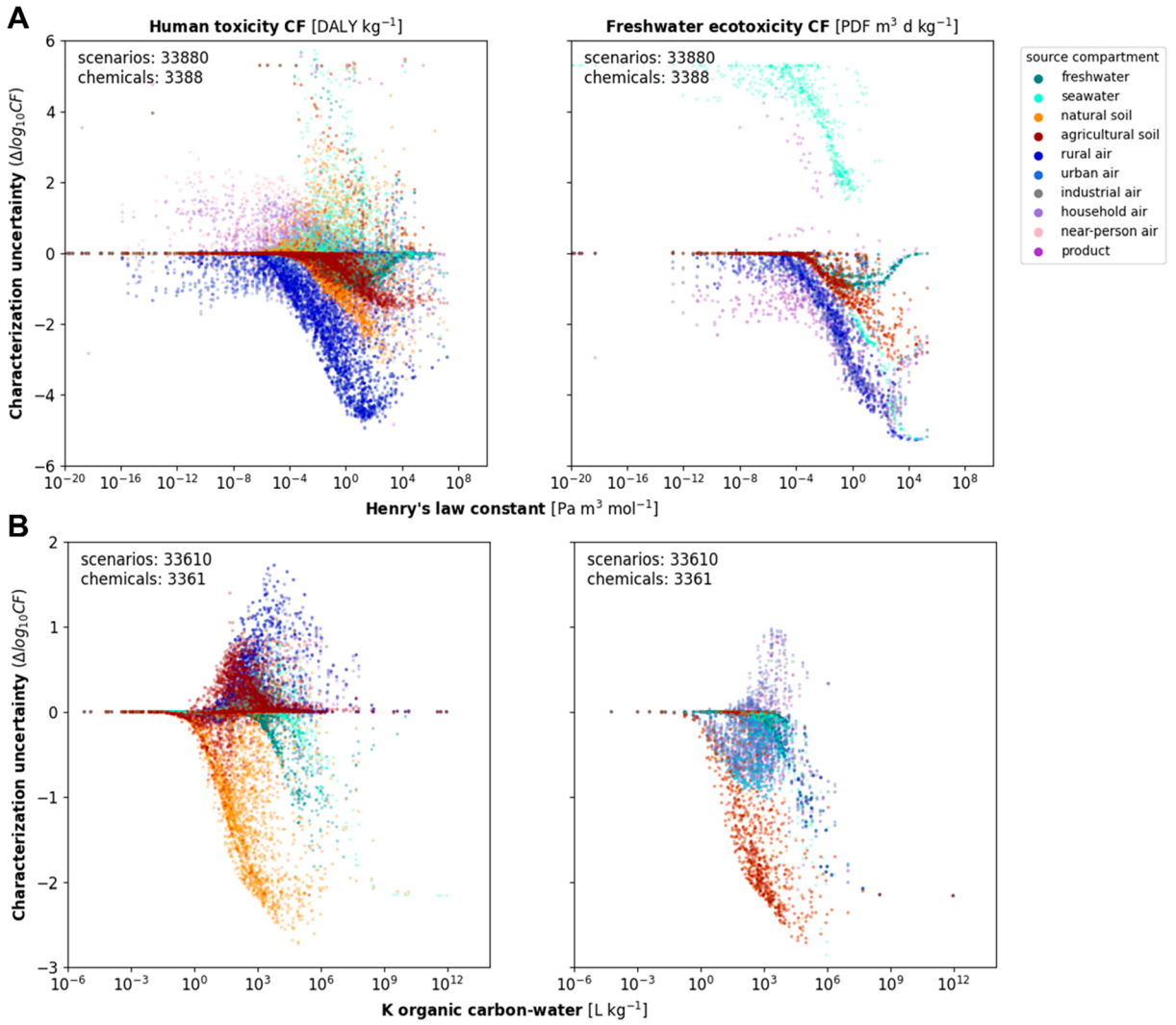


Figure 3. Uncertainty in human toxicity and freshwater ecotoxicity characterization factors (CF) as a function of (A) Henry's law constant and (B) organic carbon−water partition ratio (K organic carbon−water) across ten source compartments. DALY = disability-adjusted life years, PDF = potentially disappeared fraction of species.

uncertainty classes for vapor pressure and Henry's law constant, respectively. Characterization uncertainty was partly driven by differences in input parameter uncertainty derived from existing prediction models (GSD^2 values of 12, 39, 131, and 443 for K_{OC}, water solubility, vapor pressure, and Henry's law constant, respectively). Data availability was moderate for K_{OC} and Henry's law constants (2084 to 4492 chemicals), and high for vapor pressure and water solubility (7278 to 10300 chemicals). Our results demonstrate that these parameters have good data availability while being highly relevant for obtaining robust characterization results for many chemical-emission scenarios, making them a priority for developing ML models.

3.1.2. Environmental Degradation. The second parameter group covers environmental degradation half-lives, for which only the degradation half-life in soil ($t_{1/2}^{soil}$) was prioritized, while degradation half-lives in air ($t_{1/2}^{air}$) and water ($t_{1/2}^{water}$) were flagged as critical gaps. Notably, $t_{1/2}^{air}$ and $t_{1/2}^{water}$ reached high uncertainty classes (3.5 o 3.8 OOM), driven by high input uncertainty in $t_{1/2}^{air}$ (GSD^2 = 204), while $t_{1/2}^{water}$ gained relevance by directly influencing the exposure concentration of freshwater ecosystems. However, data availability was (very) low (419 to 749 chemicals) for all degradation half-lives except soil, which reached moderate data availability with 1774 chemicals. Higher data availability for $t_{1/2}^{soil}$ results from requirements to report this parameter for pesticide risk assessment.[53,54] Our results showed that while degradation half-lives in air and water particularly affect the robustness of characterization results, current data availability limits the potential to build ML-based prediction methods for all except degradation in soil.

3.1.3. Toxicity Effects. Toxicity effects-related parameters are the third and most important group, with 8 out of 12 parameters prioritized and the remaining four flagged as critical gaps. Uncertainty classes were high or very high (3.2 to 5.5 OOM) for eight parameters, where four parameters also showed moderate or high uncertainty classes based on the median (2.4 to 4 OOM). This general relevance across chemical-emission scenarios results from toxicity-related parameters' exclusive influence on the effect factors (EF), allowing direct input uncertainty propagation (GSD^2 ranging from 32 to 592) to the characterization results. Data availability was moderate or high for 8 out of 12 parameters, while only little data (228 to 512 chemicals) were available for four parameters related to cancer and reproductive-devel-

https://doi.org/10.1021/acs.est.3c05300
Environ. Sci. Technol. 2023, 57, 18259−18270

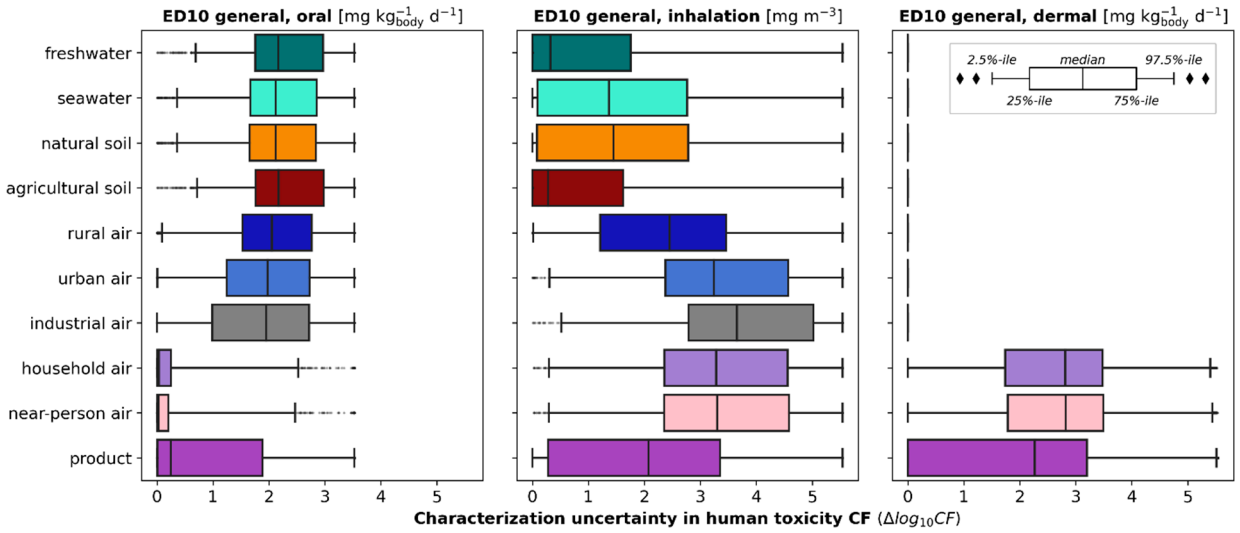


Figure 4. Distribution of uncertainty in human toxicity characterization factors (CF) attributable to oral, inhalation, and dermal exposure-related 10% response general noncancer toxicity effect doses (ED10) across ten source compartments.

opmental human toxicity via inhalation or dermal exposure. These parameters were therefore flagged as critical gaps where current data availability limits the potential to build ML-based prediction methods.

In summary, our prioritization demonstrated that parameters related to toxicity and environmental partitioning had high relevance and potential to be addressed through ML-based approaches, while the most critical gap was found for environmental degradation half-lives. Other parameters that were flagged as critical gaps were bioaccumulation factors in fish and root crops and gaseous skin permeation factors. The remaining parameters were neither prioritized for ML model development nor flagged as critical gaps, as their relevance for obtaining robust characterization results across the analyzed scenarios was low compared to the prioritized parameters.

3.2. Characterization Uncertainty Trends. Characterization uncertainty was assessed by propagating input uncertainty for 38 parameters across 3421 chemicals and ten source compartments. Input uncertainty was defined as a parameter-specific value ranging from $GSD^2 = 5$ for dissipation half-lives in plants to $GSD^2 = 864$ for gaseous skin permeation coefficients (see SI, Table S2-1). Overall uncertainty described by the median and 95th percentile of the absolute uncertainty distribution reached as high as 3 OOM ($HC20_{freshwater}$) and 5.5 OOM ($ED10^{inhalation}_{general}$), respectively (see Figure 2). Characterization uncertainties for individual chemical-emission scenarios can differ from these classes due to nonlinearity in the characterization framework. SI Figure S1-2 presents an overview of the signed characterization uncertainty distributions for all parameters, which span up to 20 orders of magnitude, with 25 parameters exhibiting positive and negative correlations across chemical-emission scenarios. Some parameters showed distinctive trends in characterization uncertainty as a function of the parameter value. As an example, Figure 3 shows the uncertainty in human and ecosystem characterization results (CF) as a function of the Henry's law constant (H) and K_{OC} with each point representing a chemical-emission scenario. While individual points were scattered, ascending and descending trends in characterization uncertainty occurred along the parameter ranges. In particular, absolute uncertainty values increased significantly for $H > 10^{-4}\ \frac{Pa \cdot m^3}{mol}$ and $K_{OC} > 10^{-1}\ \frac{L}{kg}$, where processes influenced by Henry's law constant and K_{OC} become important drivers of chemical fate.[55] Knowledge of such trends can inform ML model development to prioritize high prediction performance in critical value ranges. It also demonstrates the importance of quantifying input uncertainty at the level of individual chemicals to allow accurate propagation of uncertainty toward characterization results.

For other parameters, the characterization distribution changed significantly depending on the source compartment of the chemical emission. To illustrate, Figure 4 compares the uncertainty propagated to human toxicity characterization results from general noncancer toxicity effect doses via oral, inhalation and dermal exposure across source compartments. Characterization uncertainty attributable to oral effect doses was distributed similarly for emissions to water, soil, and far-field air compartments (mean $\cong$ 2). In contrast, inhalation effect doses were related to higher characterization uncertainty for emissions to air compartments and dermal effect doses were relevant only in near-field air and product compartments, where dermal exposure to chemicals in consumer products can be a substantial contributor to overall exposure. While our prioritization applied equal weighting across source compartments to derive a general relevance, changes in parameter relevance across source compartments are illustrated in SI Figure S1-3. These results demonstrate the importance of filling gaps for parameters with context-specific relevance to allow robust characterization across all source compartments.

Our uncertainty analysis demonstrated how input uncertainty was propagated differently for parameters and chemical-emission scenarios, highlighting the need to quantify input uncertainty at the level of individual chemicals. Some parameters were relevant across chemicals and source compartments, whereas others were particularly relevant for specific scenarios only. This demonstrated the need to eventually fill gaps for all parameters, which can significantly influence toxicity characterization results in any given scenario.

3.3. Data Availability and Chemical Space. Data availability was assessed for all 38 parameters based on our inventory of data repositories containing measured data for each parameter (see SI, Table S2-2), with numbers of unique

https://doi.org/10.1021/acs.est.3c05300
Environ. Sci. Technol. 2023, 57, 18259−18270

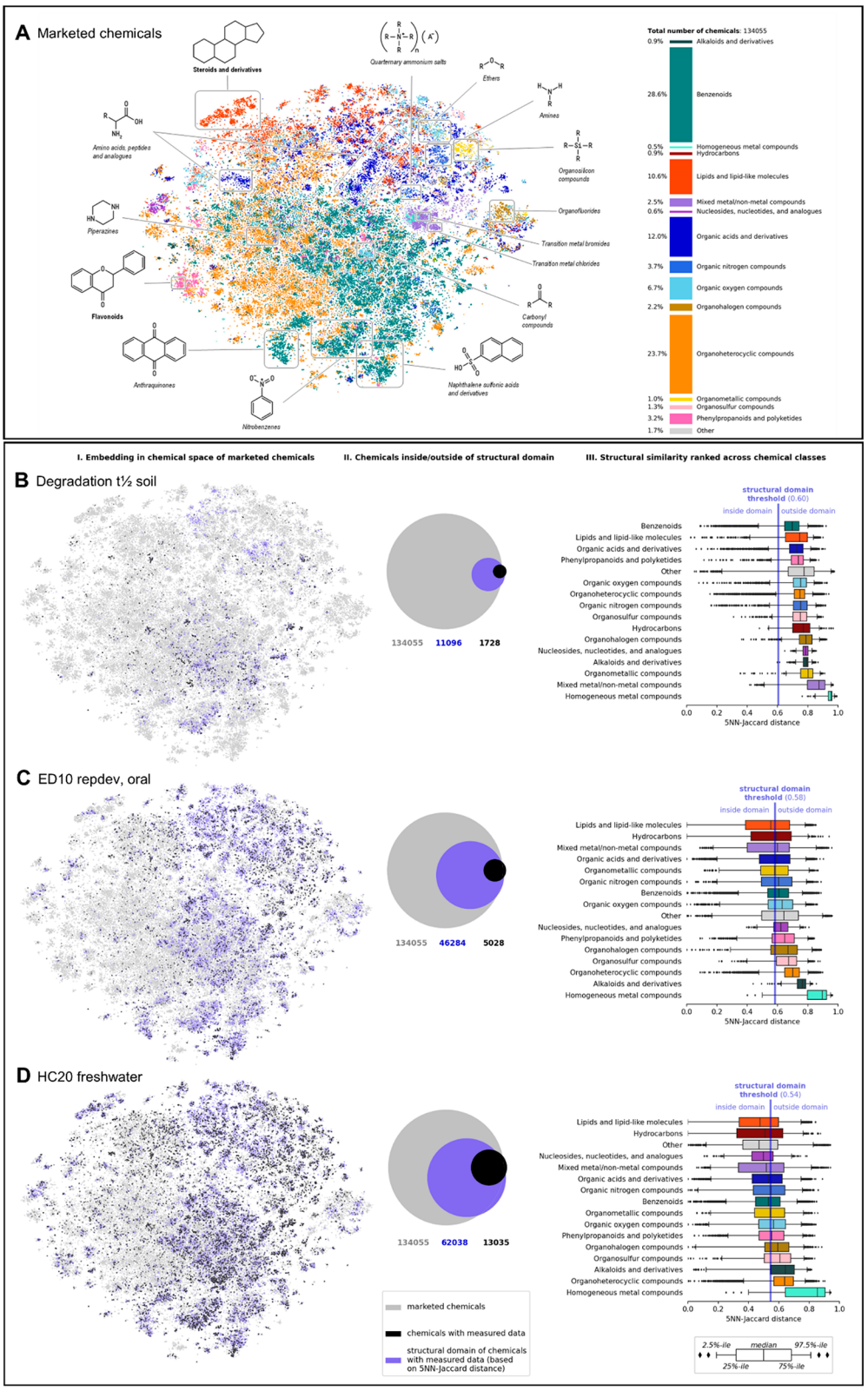


Figure 5. (A) Structural diversity of the space of marketed chemicals, colored by ClassyFire superclass with example annotations of classes (**bold**) and subclasses (*italic*) in comparison with the diversity of chemicals with measured data for three prioritized parameters, namely, (B) degradation half-life in soil (Degradation $t^1/_2$ soil), (C) 10% response effect dose for reproductive-developmental toxicity via oral exposure (ED10 repdev oral), and (D) 20% response hazardous concentration for freshwater ecotoxicity (HC20 freshwater).

chemicals ranging from <15 chemicals for gaseous skin permeation coefficients to >35 000 chemicals for octanol−water partition ratios (see Figure 2). To determine the potential of ML-based approaches to predict prioritized parameters for a wide range of chemicals, we assessed the structural diversity of chemicals with measured data relative to

https://doi.org/10.1021/acs.est.3c05300
Environ. Sci. Technol. 2023, 57, 18259−18270

134 055 marketed chemicals derived from chemical registry and use lists (see SI, Section S1.4b). Results are shown in Figure 5, presenting chemical space plot examples for three prioritized parameters (see SI, Section S1.4e for all prioritized parameters).

Figure 5A shows the two-dimensional space of marketed chemicals, with each point representing a unique chemical colored by its superclass and example annotations of classes and subclasses based on the *ClassyFire* chemical taxonomy. For better visibility, we displayed the 15 most frequent out of 27 superclasses occurring in our set of marketed chemicals, and the remaining 12 superclasses, representing 1.7% of marketed chemicals, we aggregated into "Other." Structurally similar chemicals were arranged into clusters that align well with human-defined chemical classes, illustrated by the *ClassyFire* chemical taxonomy. Organic compounds with cyclic structures dominate the chemical space, spreading over the left and bottom-right area, with benzenoids and organoheterocyclic compounds jointly making up >50% of marketed chemicals. Other chemical classes, including metal and various acyclic organic compounds, were grouped in the top-right area.

Figure 5B–D presents the chemical space of degradation half-life in soil ($t_{1/2}^{soil}$), reproductive-developmental toxicity effects via oral exposure ($ED10_{oral}^{repdev}$), and freshwater ecotoxicity effects ($HC20_{freshwater}$) in three parts: (I) The space plots show how chemicals with measured data were located across the space of marketed chemicals. (II) The Venn diagrams show the shares of chemicals with measured data and marketed chemicals inside or outside the structural domain defined by average Jaccard distances, which describe the similarity between marketed chemicals and chemicals with measured data. (III) The box plots show how average Jaccard distances ($\overline{d}_J$) are distributed across *ClassyFire* superclasses, illustrating how chemicals with measured data represent chemical classes differently. Measured $t_{1/2}^{soil}$ data were available for 1.3% of the marketed chemicals covering 8% in their structural domain, making it the least covered among prioritized parameters. Figure 5B(I) shows that chemicals with measured $t_{1/2}^{soil}$ values were sparsely distributed across the space of marketed chemicals. In contrast, measured $ED10_{oral}^{repdev}$ values were available for 4% of marketed chemicals, covering 35% in their structural domain. Chemicals with measured $ED10_{oral}^{repdev}$ spread across the space of marketed chemicals with varying local densities (see Figure 5C(I)). Thus, while chemicals with measured $t_{1/2}^{soil}$ represented all superclasses poorly with $\overline{d}_J$ = 0.69 to 0.94 and less than 25% of chemical class falling inside the structural domain, chemicals with measured $ED10_{oral}^{repdev}$ represented chemical classes to different extents. Figure 5C(III) shows that lipid and hydrocarbon classes were best represented for $ED10_{oral}^{repdev}$ with $\overline{d}_J$ = 0.51 to 0.53, while organoheterocyclic compounds, alkaloid derivatives, and homogeneous metal compounds were least represented with $\overline{d}_J$ = 0.69 to 0.83. The limited structural domain for $t_{1/2}^{soil}$ is not only a result of low data availability but also of low structural diversity. In comparison, structural domains of other small data sets, such as cancer toxicity effects via oral exposure ($TD50_{cancer}^{oral}$) and soil ecotoxicity effects ($HC20_{soil}$), covered much larger shares of 18–30% of the marketed chemicals (see SI, Section S1.4e). The importance of structural diversity can also be seen by comparing the chemical space of $ED10_{oral}^{repdev}$ and $HC20_{freshwater}$. Measured data to derive $HC20_{freshwater}$ were available for 10% of marketed chemicals with a structural domain covering 46% of marketed chemicals, making it the best-covered among prioritized parameters. However, Figure 5D(I) shows that with more than twice as many chemicals with measured data than $ED10_{oral}^{repdev}$, $HC20_{freshwater}$ covers similar parts of the space of marketed chemicals. Notably, both parameters have low representation of organoheterocyclic compounds that make up 24% of marketed chemicals, leading to a visible lack of coverage in the left part of the marketed chemical space, where most organoheterocyclic compounds are located. This lack of data cannot be explained by a lack of toxicity as seen, for example, in the curated data by Aurisano et al. (2023) for which 58% of >500 organoheterocyclic compounds with $ED10_{oral}^{repdev}$ equivalents covered in the data set fall below the median $ED10_{oral}^{repdev}$ of 16.5 $\frac{mg}{kg \cdot d}$ across all chemical classes.[56] Systematically adding measured data for organoheterocyclic compounds could therefore expand the structural domains of $ED10_{oral}^{repdev}$ and $HC20_{freshwater}$ to further improve the potential to train ML-based approaches that predict a wide range of marketed chemicals for these parameters.

Based on our results, available data for all prioritized parameters can be used to train ML models that have the potential to predict many chemicals without measured data. While measured data were available for only 1–10% of marketed chemicals, these chemicals with measured data covered 8–46% of marketed chemicals within their structural domains (see SI, Table S1-8). Our results show that the structural diversity of existing data significantly affected their capacity to represent the space of marketed chemicals. Small data sets covering diverse structures can reach equal or higher chemical space coverage compared to significantly larger data sets. All prioritized parameters covered chemicals from multiple chemical classes to varying extents. In particular, organoheterocyclic compounds, the second largest class among marketed chemicals, were insufficiently represented in all parameter data sets, leading to a significant gap in the coverage of the space of marketed chemicals. Our results demonstrated how evaluating the structural diversity of an available data set can assess the potential to train ML-based prediction models applicable to a wide range of chemicals and can also identify chemicals that could increase the potential for developing ML models that can predict the realm of marketed chemicals if measured parameter data for these chemicals were generated.

3.4. Machine Learning Potential and Limits for Addressing Data Gaps in Chemical Toxicity Characterization. Our study demonstrated the potential of ML-based approaches to address important data gaps in chemical toxicity characterization frameworks. Out of 38 input parameters, 13 were prioritized for developing ML-based approaches, notably several toxicity effect parameters. Another nine parameters, including environmental degradation half-lives, were flagged as critical data gaps. For the prioritized parameters, our results provided insights into the expected potential for making ML-based predictions across structurally diverse chemicals for the prioritized parameters. By defining the space of marketed chemicals as a reference, we created comparability across input parameters and defined a scope for the required level of predictive capacity that aligns with the broad scope of chemical toxicity assessments. Based on our results, ML-based methods have great potential to address large shares of marketed chemicals for several crucial input parameters in human toxicity and ecotoxicity characterization frameworks. In particular, vapor pressure, water solubility, and several

https://doi.org/10.1021/acs.est.3c05300
Environ. Sci. Technol. 2023, 57, 18259–18270

parameters related to human toxicity ($TD50_{oral}^{cancer}$, $ED10_{oral}^{repdev}$, $ED10_{dermal}^{general}$) and freshwater ecotoxicity ($HC20_{freshwater}$) have diverse data available that can potentially predict >30% of marketed chemicals based on our structural domain. In addition, our analysis provided insights into chemical classes that were less well represented in the parameter-specific training data. Particularly, organoheterocyclic compounds were found to be largely outside the structural domains of all prioritized parameters. Strategically targeting these chemicals underrepresented in existing parameter data sets for data curation and experimental testing could systematically broaden the chemical coverage, thereby expanding the potential of ML-based approaches to address an even wider range of marketed chemicals, while minimizing costs and effort. Our uncertainty analysis showed that it is crucial to provide chemical-specific parameter uncertainty to propagate uncertainty to toxicity characterization results due to nonlinear modeling behavior. With the increasing use of predictions alongside measured data in chemical assessments, it will become increasingly important to consider relative uncertainties from both measured and predicted data when quantifying toxicity impacts across chemicals and products. To support this aim, ML modeling approaches should be chosen that consider uncertainty and provide confidence intervals alongside each predicted value. As most ready-made ML algorithms do not innately handle uncertainty, this requires designing more advanced (combined) approaches using, for example, probabilistic modeling,[57,58] conformal prediction,[59,60] or bootstrapping approaches.[61]

Despite these promising results, the current and future potential of ML in addressing critical gaps across parameters and chemicals is constrained by several factors, the most important being data limitations. Based on current data availability, >56% of all marketed chemicals fall outside our structural domains for all prioritized parameters. Furthermore, while our identified data repositories (SI, Table S2-2) can be a suitable starting point for curating relevant training data for each parameter, it is crucial to consider the data reporting quality of data sources to create high-quality data sets for training ML models. Properly documented meta-data on measurements are required to identify high-quality data, for example, to include only data records with coherent test conditions following standardized protocols.[62] Similarly, meta-data can become important predictor variables that allow ML models to account for factors influencing parameter predictions, such as water chemistry, environmental temperature, or species characteristics. Although it is good practice to document meta-data when reporting study results[63] and the importance of providing access to meta-data is widely recognized,[64−66] they are often missing or incomplete in existing databases.[67] Similarly, many of our identified data repositories did not consistently report such information, including general testing conditions and material or species characteristics. In particular, the degradation- and toxicity-related parameters are affected by substantial differences in data quality and incomplete meta-data. High quality data available for training ML models and related structural domains are, therefore, likely lower than our provided estimate. Expanding structural domains beyond existing data or making parameters highlighted with critical data gaps suitable for ML-based approaches would require the generation of large amounts of new data, which may not be possible for all parameters due to economic and ethical constraints. In particular, parameters relying on animal tests such as toxicity-related parameters or bioaccumulation and biotransfer factors will likely be heavily restricted or replaced by new test systems such as high-throughput in vitro testing and high-content screening.[68] While their high-throughput nature makes them suitable for ML-based approaches, this new type of data will need to be linked to existing data, e.g., through IVIVE (*in vivo−in vitro* extrapolation) to seamlessly integrate into existing toxicity characterization frameworks. Future analyses could include these new parameters as well as chemical-related parameters that are relevant for other chemical assessment frameworks, such as pharmacokinetic properties for modeling internal exposure. Principle-driven or physics-informed mechanistic models may be better equipped to address data-sparse end points due to their deductive nature that allows them to extrapolate behavior more robustly beyond what is known from available data, while requiring significantly fewer data for calibration.[69] A good example of parameters that are best derived mechanistically are speciation coefficients for metal ions, which can be obtained from models replicating metal ion specific sorption and complexation reactions such as WHAM and vMinteq.[70,71] While mechanistic models provide causal explanations generally lacking from ML-based approaches, they are most suitable within well-characterized key events and driving mechanisms. Complex interactions of mechanisms may not be fully captured by a single mechanistic model or may only be described for a subset of chemical structures, which makes navigating complementary models challenging for practitioners screening a variety of chemicals. To overcome the respective limitations of mechanistic and ML modeling, both can be harnessed to synergize scientific understanding and available (old and new) data, for example, in the form of physics-based ML[72] or surrogate modeling.[69,73] The development of such approaches will require increased efforts and resources to address marketed chemicals that are currently outside the intrinsic domain of ML models based on available data, as illustrated by our structural domains. Thus, while ML models can make a significant contribution to filling data gaps in chemical toxicity characterization frameworks, more data and advanced approaches will be needed to allow robust toxicity characterization for all marketed chemicals. Strategic integration of (hybrid) model development and (novel) data generation efforts will be crucial to increase chemical coverage and reliability of toxicity characterization results for use in LCIA, chemical substitution, risk screening, and SSbD to support decision-making in regulatory and other chemical management frameworks.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.3c05300.

Additional details on applied methods including underlying literature searches and modeling choices, and additional results figures for all input parameters (PDF)

Tables containing identified GSD^2 values and data repositories (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Kerstin von Borries − *Quantitative Sustainability Assessment, Department of Environmental and Resource Engineering,*

https://doi.org/10.1021/acs.est.3c05300
Environ. Sci. Technol. 2023, 57, 18259−18270

Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; orcid.org/0000-0001-9438-6562; Email: kejbo@dtu.dk

Peter Fantke − *Quantitative Sustainability Assessment, Department of Environmental and Resource Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark;* orcid.org/0000-0001-7148-6982; Email: pefan@dtu.dk

Authors

Hanna Holmquist − *IVL Swedish Environmental Research Institute, 411 33 Göteborg, Sweden*

Marissa Kosnik − *Quantitative Sustainability Assessment, Department of Environmental and Resource Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark;* Present Address: (M.K.) Department of Environmental Toxicology, Swiss Federal Institute of Aquatic Science and Technology, EAWAG, 8600 Dübendorf, Switzerland

Katie V. Beckwith − *Centre for Molecular Informatics, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom*

Olivier Jolliet − *Quantitative Sustainability Assessment, Department of Environmental and Resource Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*

Jonathan M. Goodman − *Centre for Molecular Informatics, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom;* orcid.org/0000-0002-8693-9136

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.3c05300

Author Contributions

Kerstin von Borries: Conceptualization, Methodology, Literature review, Data curation, Modeling, Formal analysis, Visualization, Writing−original draft; Hanna Holmquist: Conceptualization, Methodology, Supervision, Writing−review and editing; Marissa Kosnik: Conceptualization, Methodology, Supervision, Writing−review and editing; Katie Beckwith: Conceptualization, Writing−review and editing; Olivier Jolliet: Validation, Writing - review and editing; Jonathan Goodman: Conceptualization, Writing−review and editing; Peter Fantke: Resources, Conceptualization, Methodology, Supervision, Validation, Literature review, Writing−review and editing

Funding

This work was financially supported by the "Safe and Efficient Chemistry by Design (SafeChem)" project funded by the Swedish Foundation for Strategic Environmental Research (Grant No. DIA 2018/11) and by the PARC project (Grant No. 101057014) funded under the European Union's Horizon Europe Research and Innovation program.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Kamel Mansouri (NIEHS National Institute of Environmental Health Sciences) for providing further information on the OPERA models, Kathrin Fenner (EAWAG) for providing access to chemicals with measured data in the EAWAG-Soil database, Susan Oginah (Technical University of Denmark) for support in working with the SOLUTIONS ecotoxicity database, David Wishart and Siyang Tian (University of Alberta) for support in applying the *ClassyFire* chemical taxonomy, Mikolaj Owsianiak (Technical University of Denmark) for input on speciation coefficients, and Philip von Borries for his graphical design support.

■ REFERENCES

(1) UNEP. *Global Chemicals Outlook II From Legacies to Innovative Solutions: Implementing the 2030 Agenda for Sustainable Development-Synthesis Report*; 2019.

(2) Persson, L.; Carney Almroth, B. M.; Collins, C. D.; Cornell, S.; de Wit, C. A.; Diamond, M. L.; Fantke, P.; Hassellöv, M.; MacLeod, M.; Ryberg, M. W.; Søgaard Jørgensen, P.; Villarrubia-Gómez, P.; Wang, Z.; Hauschild, M. Z. Outside the Safe Operating Space of the Planetary Boundary for Novel Entities. *Environ. Sci. Technol.* **2022**, *56*, No. 1510.

(3) Grundmann, V.; Bilitewski, B.; Zehm, A.; Darbra, R. M.; Barceló, D. Risk-Based Management of Chemicals and Products in a Circular Economy at a Global Scale-Impacts of the FP7 Funded Project RISKCYCLE. *Environmental Sciences Europe* **2013**, *25*, 14.

(4) US EPA. *Science Policy Council Handbook: Risk Characterization*. 2000. www.epa.gov (accessed 2023-01-30).

(5) European Commission. *European Union System for the Evaluation of Substances 2.0 (EUSES 2.0)*; 2004. by the National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands (RIVM Report no. 601900005). Available via the European Chemicals Bureau, https://www.echa.europa.eu/documents/10162/28293536/euses_2-1_background_document_en.pdf/51d73868-784b-1476-f201-8a4ba17e78e0 (accessed 2023-01-30).

(6) Shin, H. M.; Ernstoff, A.; Arnot, J. A.; Wetmore, B. A.; Csiszar, S. A.; Fantke, P.; Zhang, X.; McKone, T. E.; Jolliet, O.; Bennett, D. H. Risk-Based High-Throughput Chemical Screening and Prioritization Using Exposure Models and in Vitro Bioactivity Assays. *Environ. Sci. Technol.* **2015**, *49* (11), 6760−6771.

(7) Fantke, P.; Chiu, W. A.; Aylward, L.; Judson, R.; Huang, L.; Jang, S.; Gouin, T.; Rhomberg, L.; Aurisano, N.; McKone, T.; Jolliet, O. Exposure and Toxicity Characterization of Chemical Emissions and Chemicals in Products: Global Recommendations and Implementation in USEtox. *International Journal of Life Cycle Assessment* **2021**, *26* (5), 899−915.

(8) Owsianiak, M.; Hauschild, M. Z.; Posthuma, L.; Saouter, E.; Vijver, M. G.; Backhaus, T.; Douziech, M.; Schlekat, T.; Fantke, P. Ecotoxicity Characterization of Chemicals: Global Recommendations and Implementation in USEtox. *Chemosphere* **2023**, *310*, 136807.

(9) Sala, S.; Biganzoli, F.; Mengual, E. S.; Saouter, E. Toxicity Impacts in the Environmental Footprint Method: Calculation Principles. *International Journal of Life Cycle Assessment* **2022**, *27* (4), 587−602.

(10) Sala, S.; Goralczyk, M. Chemical Footprint: A Methodological Framework for Bridging Life Cycle Assessment and Planetary Boundaries for Chemical Pollution. *Integr Environ. Assess Manag* **2013**, *9* (4), 623−632.

(11) Fantke, P.; Huang, L.; Overcash, M.; Griffing, E.; Jolliet, O. Life Cycle Based Alternatives Assessment (LCAA) for Chemical Substitution. *Green Chem.* **2020**, *22* (18), 6008−6024.

(12) Kosnik, M. B.; Hauschild, M. Z.; Fantke, P. Toward Assessing Absolute Environmental Sustainability of Chemical Pollution. *Environ. Sci. Technol.* **2022**, *56*, 4776−4787.

(13) Fantke, P.; Illner, N. Goods That Are Good Enough: Introducing an Absolute Sustainability Perspective for Managing Chemicals in Consumer Products. *Current Opinion in Green and Sustainable Chemistry.* **2019**, *15*, 91−97.

(14) Directorate-General for Research and Innovation (European Commission). *Strategic Research and Innovation Plan for Safe and Sustainable Chemicals and Materials*; 2022; DOI: 10.2777/876851.

(15) Wang, Z.; Walker, G. W.; Muir, D. C. G.; Nagatani-Yoshida, K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environ. Sci. Technol.* **2020**, *54* (5), 2575−2584.
(16) Kosnik, M. B.; Kephalopoulos, S.; Muñoz, A.; Aurisano, N.; Cusinato, A.; Dimitroulopoulou, S.; Slobodnik, J.; De Mello, J.; Zare Jeddi, M.; Cascio, C.; Ahrens, A.; Bruinen de Bruin, Y.; Lieck, L.; Fantke, P. Advancing Exposure Data Analytics and Repositories as Part of the European Exposure Science Strategy 2020−2030. *Environ. Int.* **2022**, *170*, 107610.
(17) Fantke, P.; Cinquemani, C.; Yaseneva, P.; De Mello, J.; Schwabe, H.; Ebeling, B.; Lapkin, A. A. Transition to Sustainable Chemistry through Digitalization. *Chem.* **2021**, *7*, 2866−2882.
(18) Fantke, P.; Aurisano, N.; Provoost, J.; Karamertzanis, P. G.; Hauschild, M. Toward Effective Use of REACH Data for Science and Policy. *Environment International* **2020**, *135*, 105336.
(19) Ingre-Khans, E.; Ågerstrand, M.; Beronius, A.; Rudén, C. Transparency of Chemical Risk Assessment Data under REACH. *Environ. Sci. Process Impacts* **2016**, *18* (12), 1508−1518.
(20) Gilbert, N. Legal Tussle Delays Launch of Huge Toxicity Database. *Nature* **2016**, DOI: 10.1038/nature.2016.19365.
(21) Mahalakshmi, P. S.; Jahnavi, Y. A REVIEW ON QSAR STUDIES. *International Journal of Advances in Pharmacy and Biotechnology* **2020**, *6* (2), 19−23.
(22) Quintero, F. A.; Patel, S. J.; Muñoz, F.; Sam Mannan, M. Review of Existing QSAR/QSPR Models Developed for Properties Used in Hazardous Chemicals Classification System. *Ind. Eng. Chem. Res.* **2012**, *51* (49), 16101−16115.
(23) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'Min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.
(24) Wambaugh, J. F.; Bare, J. C.; Carignan, C. C.; Dionisio, K. L.; Dodson, R. E.; Jolliet, O.; Liu, X.; Meyer, D. E.; Newton, S. R.; Phillips, K. A.; Price, P. S.; Ring, C. L.; Shin, H. M.; Sobus, J. R.; Tal, T.; Ulrich, E. M.; Vallero, D. A.; Wetmore, B. A.; Isaacs, K. K. New Approach Methodologies for Exposure Science. *Current Opinion in Toxicology.* **2019**, *15*, 76−92.
(25) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J. Cheminform* **2018**, *10* (1), 10 DOI: 10.1186/s13321-018-0263-1.
(26) U.S. EPA. *User's Guide for T. E. S. T. (Toxicity Estimation Software Tool) Version 5.1 A Java Application to Estimate Toxicities and Physical Properties from Molecular Structure*; 2020. https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test (accessed 2021-12-16).
(27) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-Source QSAR Models for PKa Prediction Using Multiple Machine Learning Approaches. *J. Cheminform* **2019**, *11* (1), 60 DOI: 10.1186/s13321-019-0384-1.
(28) Xiong, J.; Li, Z.; Wang, G.; Fu, Z.; Zhong, F.; Xu, T.; Liu, X.; Huang, Z.; Liu, X.; Chen, K.; Jiang, H.; Zheng, M. Multi-Instance Learning of Graph Neural Networks for Aqueous PKa Prediction. *Bioinformatics* **2022**, *38* (3), 792−798.
(29) Gao, F.; Shen, Y.; Brett Sallach, J.; Li, H.; Zhang, W.; Li, Y.; Liu, C. Predicting Crop Root Concentration Factors of Organic Contaminants with Machine Learning Models. *J. Hazard Mater.* **2022**, *424*, No. 127437.
(30) Gao, F.; Shen, Y.; Sallach, J. B.; Li, H.; Liu, C.; Li, Y. Direct Prediction of Bioaccumulation of Organic Contaminants in Plant Roots from Soils with Machine Learning Models Based on Molecular Structures. *Environ. Sci. Technol.* **2021**, *55*, 16358.
(31) Bagheri, M.; Al-jabery, K.; Wunsch, D.; Burken, J. G. Examining Plant Uptake and Translocation of Emerging Contaminants Using Machine Learning: Implications to Food Security. *Sci. Total Environ.* **2020**, *698*, 133999.
(32) Hou, P.; Jolliet, O.; Zhu, J.; Xu, M. Estimate Ecotoxicity Characterization Factors for Chemicals in Life Cycle Assessment Using Machine Learning Models. *Environ. Int.* **2020**, *135*, 105393.
(33) Hou, P.; Zhao, B.; Jolliet, O.; Zhu, J.; Wang, P.; Xu, M. Rapid Prediction of Chemical Ecotoxicity through Genetic Algorithm Optimized Neural Network Models. *ACS Sustain Chem. Eng.* **2020**, *8* (32), 12168−12176.
(34) Gao, F.; Zhang, W.; Baccarelli, A. A.; Shen, Y. Predicting Chemical Ecotoxicity by Learning Latent Space Chemical Representations. *Environ. Int.* **2022**, *163*, 107224.
(35) Song, R.; Li, D.; Chang, A.; Tao, M.; Qin, Y.; Keller, A. A.; Suh, S. Accelerating the Pace of Ecotoxicological Assessment Using Artificial Intelligence. *Ambio* **2022**, *51*, 598.
(36) Hiki, K.; Iwasaki, Y. Can We Reasonably Predict Chronic Species Sensitivity Distributions from Acute Species Sensitivity Distributions? *Environ. Sci. Technol.* **2020**, *54* (20), 13131−13136.
(37) Iwasaki, Y.; Sorgog, K. Estimating Species Sensitivity Distributions on the Basis of Readily Obtainable Descriptors and Toxicity Data for Three Species of Algae, Crustaceans, and Fish. *PeerJ.* **2021**, *9*, e10981.
(38) Rosenbaum, R. K.; Bachmann, T. M.; Gold, L. S.; Huijbregts, M. A. J.; Jolliet, O.; Juraske, R.; Koehler, A.; Larsen, H. F.; MacLeod, M.; Margni, M.; McKone, T. E.; Payet, J.; Schuhmacher, M.; Van De Meent, D.; Hauschild, M. Z. USEtox - The UNEP-SETAC Toxicity Model: Recommended Characterisation Factors for Human Toxicity and Freshwater Ecotoxicity in Life Cycle Impact Assessment. *International Journal of Life Cycle Assessment* **2008**, *13* (7), 532−546.
(39) Westh, T. B.; Hauschild, M. Z.; Birkved, M.; Jørgensen, M. S.; Rosenbaum, R. K.; Fantke, P. The USEtox Story: A Survey of Model Developer Visions and User Requirements. *International Journal of Life Cycle Assessment* **2015**, *20* (2), 299−310.
(40) Gentil, C.; Basset-Mens, C.; Manteaux, S.; Mottes, C.; Maillard, E.; Biard, Y.; Fantke, P. Coupling Pesticide Emission and Toxicity Characterization Models for LCA: Application to Open-Field Tomato Production in Martinique. *J. Clean Prod* **2020**, *277*, 124099.
(41) Huang, L.; Fantke, P.; Ritscher, A.; Jolliet, O. Chemicals of Concern in Building Materials: A High-Throughput Screening. *J. Hazard Mater.* **2022**, *424*, 127574.
(42) Aurisano, N.; Huang, L.; Milà i Canals, L.; Jolliet, O.; Fantke, P. Chemicals of Concern in Plastic Toys. *Environ. Int.* **2021**, *146*, 106194.
(43) Gandhi, N.; Diamond, M. L.; Van De Meent, D.; Huijbregts, M. A. J.; Peijnenburg, W. J. G. M.; Guinée, J. New Method for Calculating Comparative Toxicity Potential of Cationic Metals in Freshwater: Application to Copper, Nickel, and Zinc. *Environ. Sci. Technol.* **2010**, *44* (13), 5195−5201.
(44) Artrith, N.; Butler, K. T.; Coudert, F. X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nature Chemistry.* **2021**, *13*, 505−508.
(45) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: A Community Data Resource for Environmental Chemistry. *J. Cheminform* **2017**, *9* (1), 61 DOI: 10.1186/s13321-017-0247-6.
(46) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *J. Cheminform* **2016**, *8* (1), 1−20.
(47) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf Model* **2010**, *50* (5), 742−754.
(48) RDKit. *RDKit: Open-Source Cheminformatics*, 2022, DOI: 10.5281/zenodo.6961488.
(49) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579−2605.
(50) Poličar, P. *openTSNE*. https://opentsne.readthedocs.io/en/latest/index.html (accessed 2022-12-15).

(51) Jones, C. W.; Lawal, W.; Xu, X. Emerging Chemistry & Machine Learning. *JACS Au* **2022**, *2* (3), 541−542.
(52) van Rossum, G.; the Python development team. *The Python Language Reference Release 3.10.11*; 2023. https://www.python.org/downloads/release/python-3100/ (accessed 2023-06-13).
(53) U.S. EPA. *Data Requirements for Pesticides*; U. S. Environmental Protection Agency, 2023. https://www.ecfr.gov/current/title-40/chapter-I/subchapter-E/part-158 (accessed 2023-02-13).
(54) EC. *Regulation Setting out the Data Requirements for Active Substances*; European Commission: 2013.
(55) Henderson, A. D.; Hauschild, M. Z.; Van De Meent, D.; Huijbregts, M. A. J.; Larsen, H. F.; Margni, M.; McKone, T. E.; Payet, J.; Rosenbaum, R. K.; Jolliet, O. USEtox Fate and Ecotoxicity Factors for Comparative Assessment of Toxic Emissions in Life Cycle Analysis: Sensitivity to Key Chemical Properties. *International Journal of Life Cycle Assessment* **2011**, *16* (8), 701−709.
(56) Aurisano, N.; Jolliet, O.; Chiu, W. A.; Judson, R.; Jang, S.; Unnikrishnan, A.; Kosnik, M. B.; Fantke, P. Probabilistic Points of Departure and Reference Doses for Characterizing Human Non-cancer and Developmental/Reproductive Effects for 10,145 Chemicals. *Environ. Health Perspect* **2023**, *131* (3), 37016.
(57) Mervin, L. H.; Trapotsi, M. A.; Afzal, A. M.; Barrett, I. P.; Bender, A.; Engkvist, O. Probabilistic Random Forest Improves Bioactivity Predictions Close to the Classification Threshold by Taking into Account Experimental Uncertainty. *J. Cheminform* **2021**, *13* (1), 62 DOI: 10.1186/s13321-021-00539-7.
(58) Allen, T. E. H.; Middleton, A. M.; Goodman, J. M.; Russell, P. J.; Kukic, P.; Gutsell, S. Towards Quantifying the Uncertainty in in Silico Predictions Using Bayesian Learning. *Computational Toxicology* **2022**, *23*, 100228.
(59) Svensson, F.; Aniceto, N.; Norinder, U.; Cortes-Ciriano, I.; Spjuth, O.; Carlsson, L.; Bender, A. Conformal Regression for Quantitative Structure-Activity Relationship Modeling - Quantifying Prediction Uncertainty. *J. Chem. Inf Model* **2018**, *58* (5), 1132−1140.
(60) Zhang, J.; Norinder, U.; Svensson, F. Deep Learning-Based Conformal Prediction of Toxicity. *J. Chem. Inf Model* **2021**, *61* (6), 2648−2657.
(61) Pradeep, P.; Paul Friedman, K.; Judson, R. Structure-Based QSAR Models to Predict Repeat Dose Toxicity Points of Departure. *Computational Toxicology* **2020**, *16*, 100139.
(62) Fantke, P.; Arnot, J. A.; Doucette, W. J. Improving Plant Bioaccumulation Science through Consistent Reporting of Experimental Data. *Journal of Environmental Management.* **2016**, *181*, 374−384.
(63) OECD. *OECD Guidelines for the Testing of Chemicals*; 2023.
(64) Olker, J. H.; Elonen, C. M.; Pilli, A.; Anderson, A.; Kinziger, B.; Erickson, S.; Skopinski, M.; Pomplun, A.; LaLone, C. A.; Russom, C. L.; Hoff, D. The ECOTOXicology Knowledgebase: A Curated Database of Ecologically Relevant Toxicity Tests to Support Environmental Research and Risk Assessment. *Environ. Toxicol. Chem.* **2022**, *41* (6), 1520−1539.
(65) Stepanov, D.; Canipa, S.; Wolber, G. HuskinDB, a Database for Skin Permeation of Xenobiotics. *Sci. Data* **2020**, *7* (1), 426.
(66) Latino, D. A. R. S.; Wicker, J.; Gütlein, M.; Schmid, E.; Kramer, S.; Fenner, K. Eawag-Soil in EnviPath: A New Resource for Exploring Regulatory Pesticide Soil Biodegradation Pathways and Half-Life Data. *Environ. Sci. Process Impacts* **2017**, *19* (3), 449−464.
(67) Hrovat, M.; Segner, H.; Jeram, S. Variability of in Vivo Fish Acute Toxicity Data. *Regul. Toxicol. Pharmacol.* **2009**, *54* (3), 294−300.
(68) Stucki, A. O.; Barton-Maclaren, T. S.; Bhuller, Y.; Henriquez, J. E.; Henry, T. R.; Hirn, C.; Miller-Holt, J.; Nagy, E. G.; Perron, M. M.; Ratzlaff, D. E.; Stedeford, T. J.; Clippinger, A. J. Use of New Approach Methodologies (NAMs) to Meet Regulatory Requirements for the Assessment of Industrial Chemicals and Pesticides for Effects on Human Health. *Frontiers in Toxicology*. **2022**, DOI: 10.3389/ftox.2022.964553.
(69) Baker, R. E.; Peña, J. M.; Jayamohan, J.; Jérusalem, A. Mechanistic Models versus Machine Learning, a Fight Worth Fighting for the Biological Community? *Biol. Lett.* **2018**, *14* (5), 20170660.
(70) Gustafsson, J. P. *Visual MINTEQ*. https://vminteq.com/ (accessed 2023-04-28).
(71) UKCEH. *Windermere Humic Aqueous Model (WHAM)*; UK Centre for Ecology and Hydrology: 2023. https://www.ceh.ac.uk/services/windermere-humic-aqueous-model-wham (accessed 2023-04-28).
(72) Vadyala, S. R.; Betgeri, S. N.; Matthews, J. C.; Matthews, E. A Review of Physics-Based Machine Learning in Civil Engineering. *Results in Engineering* **2022**, *13*, 100316.
(73) Fedik, N.; Zubatyuk, R.; Kulichenko, M.; Lubbers, N.; Smith, J. S.; Nebgen, B.; Messerly, R.; Li, Y. W.; Boldyrev, A. I.; Barros, K.; Isayev, O.; Tretiak, S. Extending Machine Learning beyond Interatomic Potentials for Predicting Molecular Properties. *Nature Reviews Chemistry*. **2022**, *6*, 653−672.