UNIVERSITY OF CALIFORNIA

Santa Barbara

Machine Learning for Addressing Data Deficiencies in Life Cycle Assessment

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Environmental Science and Management

by

Runsheng Song

Committee in charge:

Professor Sangwon Suh, Co-Chair

Professor Arturo A. Keller, Co-Chair

Professor Krzysztof Janowicz

March 2019

The dissertation of Runsheng Song is approved.

_____

Arturo A. Keller

_____

Krzysztof Janowicz

_____

Sangwon Suh, Committee Chair

March 2019

Machine Learning for Addressing Data Deficiencies in Life Cycle Assessment

by

Runsheng Song

ACKNOWLEDGEMENTS

I have the privilege to work with a group of smart people during my stay at the Bren School of Environmental Science and Management, UC Santa Barbara. I would like to first give my thanks to Prof. Sangwon Suh, my PhD advisor, who provides me countless guide, and offered invaluable helps during many difficult times over my PhD career. I would like to thank Prof. Arturo A. Keller, who is always encouraging and generous when I was confused or making mistake over the development of my research. I also want to thank Prof. Krzysztof Janowicz, who shines his wisdom from another field to my PhD research, and always pointing out the room for improvement in my study which I could never figure out without him. For what I learnt from my professors, I appreciate that. I want to give thanks to my lab mates, who shared probably one of the most important periods in my career. I learnt how to be humble, kind, patient and professional working with such a smart group of people. And after all, although the group might sometime disagree with each other in the CLiCC project, we still made a great achievement by building a tool that could help the other researchers in their study. For what we achieved together, I appreciate that. I want to thank my parents, although who are far away from me when I am studying abroad. They offered encouragement and helps to me without looking for payback. And I always know that I have a harbor to return if the waves are too strong. For what I received from them, I appreciate that. PhD is never an easy journey, although I am just starting my professional career, but I shall never forget what the difficult time taught me during this journey: the best thing is yet to come, and don't give up

12/31/2018,

Beijing

VITA OF RUNSHENG SONG

Dec 2018

EDUCATION

- Bachelor of Engineering, Civil & Environmental Engineering, Huazhong University of Science and Technology, China, July 2012
- Master of Environmental Science and Management, University of California, Santa Barbara, June 2014
- Doctor of Philosophy in Environmental Science and Technology, University of California, Santa Barbara, March 2019 (expected)

PUBLICATIONS

1. ***Song, R.***, *Qin, Y., Suh, S., & Keller, A. A. (2017). Dynamic model for the stocks and release flows of engineered nanomaterials. Environmental science & technology, 51(21), 12424-12433.*

2. *Palazzo, J., Liu, O. R., Stillinger, T.,* ***Song, R.****, Wang, Y., Hiroyasu, E. H., ... & Tague, C. (2017). Urban responses to restrictive conservation policy during drought. Water Resources Research, 53(5), 4459-4475.*

3. ***Song, R.****, Keller, A. A., & Suh, S. (2017). Rapid life-cycle impact screening using artificial neural networks. Environmental science & technology, 51(18), 10777-10785.*

4. *Tao, M., Li, D.,* ***Song, R.****, Suh, S., & Keller, A. A. (2018). OrganoRelease–A framework for modeling the release of organic chemicals from the use and post-use of consumer products. Environmental Pollution, 234, 751-761.*

AWARDS

- Decker's Fellowship (2015, 2016, 2017)

ACADEMIC ACTIVITIES

- 2013, Researcher, National Center for Ecological Analysis and Synthesis, Santa Barbara
- 2012 – 2014 Researcher, Group Project, Bren School of Environmental Science and Management, University of California, Santa Barbara
- 2014 – 2018 Graduate Research Assistant, Chemical Life Cycle Collaborative, University of California, Santa Barbara
- 2015 – 2018 Reviewer for *Environmental Science & Technology* and *Journal of Nanoparticle Research*

TALKS AT CONFERENCES

1. American Center for Life Cycle Assessment (ACLCA) 2015, Vancouver, CA "*Estimating Intermediate and Elementary Flow in Chemical Manufacturing using Artificial Neural Networks*"
2. American Center for Life Cycle Assessment (ACLCA) 2016, Charleston, US "*Strategy for Developing Gate-to-Gate Chemical Life Cycle Inventory*"
3. Computational LCA Workshop 2017, Zurich, CH (Joined from Remote) "*Rapid Life-Cycle Impact Screening for Decision-Support at Early Stage Chemical Design*"
4. The International Society for Industrial Ecology 2017, Chicago, US Workshop for Chemical Life Cycle Collaboration (CLiCC): "*Life Cycle Inventory Module*" and "*Life Cycle Impact Assessment Module*"
5. The 8th International Conference on Life Cycle Management 2017, Luxembourg, Luxembourg Workshop for Chemical Life Cycle Collaboration (CLiCC): "*Life Cycle Inventory Module*"
6. American Center for Life Cycle Assessment (ACLCA) 2018, Fort Collins, US Workshop for Chemical Life Cycle Collaboration (CLiCC): "*Life Cycle Inventory Module*" and "*Species Sensitivity Distributions for Organic Chemicals Using Artificial Neural Networks*"

TALKS AT SEMINARS

1. PhD Symposium of Bren School 2016, Santa Barbara, US. "*Rapid chemical life-cycle impact screening for decision-support at early design stage*"
2. CLiCC Student Workshop 2016, Santa Barbara, US. "*The Use Predictive Tools and Machine Learning in CLiCC Project*"
3. CLiCC Annual Review Meeting 2015 – 2018, Santa Barbara, US. Multiple Topics

ABSTRACT

Machine Learning for Addressing Data Deficiencies in Life Cycle Assessment

by

Runsheng Song

Life Cycle Assessment (LCA) is a tool that can be used to assess the impacts of chemicals over the entire life cycle. As the large number of new chemicals being invented every day, the costs and time needed to collect necessary data for LCA studies pose a challenge to LCA practitioners, as the speed of LCA studies cannot keep up with the speed of new chemical development. In practice, therefore, LCAs are conducted in the presence of data gaps and proxy values, limiting the relevance and quality of the results. As the techniques of machine learning evolves, a new opportunity to improve on data deficiencies and on the quality of LCA emerged. This dissertation is an attempt to harness the power of machine learning techniques to address the data deficiencies in LCA. It consists of four chapters: (1) Introduction. (2) Rapid life-cycle impact screening for decision-support using artificial neural networks. (3) Species Sensitivity Distributions Derived for a Large Number of Chemicals Using Artificial Neural Networks. IV. (4) Reducing the Uncertainty of the Characterization Factors in *USEtox* by Machine Learning – A Case Study for Aquatic Ecotoxicity. Each chapter is elaborated briefly below.

The first chapter is the general introduction. The second chapter aims to demonstrate the method of estimating the characterized results using Artificial Neural Networks (ANNs). Due to the lack of necessary data, very limited amount of

characterized results for organic chemicals exist. In this chapter, I developed ANNs to estimate the characterized results of chemicals. Using molecular structure information as an input, I trained multilayer ANNs for the characterized results of chemicals on six impact categories: (1) global warming. (2) acidification. (3) cumulative energy demand. (4) human health. (5) ecosystem quality. (6) eco-indicator 99. The application domain (AD) of the model was estimated for each impact category within which the model exhibits higher reliability. As a result, the ANN models for acidification, human health, and eco-indicator 99 showed relatively higher performances with R^2 values of 0.73, 0.71, and 0.87, respectively. This chapter indicates that ANN models can serve as an initial screening tool for estimating life-cycle impacts of chemicals for certain impact categories in the absence of more reliable information.

The second chapter aims to estimate the ecotoxicological impact of chemicals using machine learning models. In chemical impact assessment, the overall ecotoxicological impact of a chemical to ecosystem, also known as the Effect Factor (EFs), is derived from the toxicity to multiple species through Species Sensitivity Distribution (SSDs). In the third chapter, I turned to estimate the chemical toxicities to several aquatic species with machine learning models, and then use them to build SSD, and to estimate the EF of organic chemicals. Over 2,000 experimental toxicity data were collected for 8 aquatic species from 20 sources, and an ANN model for each of the species was trained to estimate the Lethal Concentration (LC50) based on molecular structure. The 8 ANN models showed R^2 scores of 0.54 to 0.75 (average 0.67, medium 0.69) on testing data. The toxicity values predicted by the ANN models were then used to

fit SSDs using bootstrapping method. At the end, the models were applied to generate SSDs for 8,424 chemicals in the ToX21 database.

The last chapter of this dissertation aims to reduce the uncertainty of an existing chemical fate model using machine learning techniques. Fate Factor (FF), which accounts the persistence of chemicals in environmental compartments, is an intermediate input in to calculate the characterized results of life cycle impact assessment. The most widely used tool to calculate chemical FFs: *USEtox*, requires several chemical properties as inputs, including: octanol-water partitioning coefficient (K_{ow}) and vapor pressure at 25 ℃ (P_{vap25}). When those chemical properties are missing, *USEtox* provides proxy methods to estimate them. In the fourth chapter, I seek to answer the question that whether replacing the current proxy methods with machine learning models are always improving the accuracy of FFs. The contribution of each chemical property to the FFs was evaluated. And ANN-based predictive models were developed to predict these chemical properties. The uncertainty of the current proxy methods in the *USEtox*'s FF model and the newly developed ANN models were compared. New FFs for the chemicals in the ToX21 database were calculated using the best predictive model when experimental properties were unknown. The EFs generated by the models in the second chapter were estimated. Lastly, more than 300 new CFs with good prediction confidence for the organic chemicals in the ToX21 database were calculated. These CFs are new to the field of LCA and can be used to reduce the uncertainty of LCA studies when the measured data isn't available.

I. Introduction

A. Background

Life Cycle Assessment. Life Cycle Assessment (LCA) is a tool to assess the environmental and human health impacts of product throughout its life cycle[1]. A typical LCA study consists of four phases: Goal and Scope Definition, Life Cycle Inventory Analysis (LCI), Life Cycle Impact Assessment (LCIA), and Interpretation[2]. In the past decades, LCA has gained its importance in many areas. For example, the eco-labeling program: Environmental Product Declaration (EPD), and the Leadership in Energy and Environmental Design (LEED) program, report LCA data[3]. ISO-compliant LCA studies have also become a standard methods to report the sustainability of top companies, like Coca-Cola, Dow Chemical and DuPont[4,5]. In literatures, the methodologies of LCA are prevalent. For example, the handbook of LCA by Guinee et al, the computational guide of LCA by Heijungs and Suh, the overview of LCAs in the past decade By Finnveden et al. have been cited more than five thousand times in total[1,6,7].

The stage of LCIA is an important step that converts the mass of emission to the scores which reflects the environmental or human health impact. The conversion factor from emission to impact is the so called "Characterization Factor" (CF). CF is associated with chemical emission in different environmental compartment (i.e., water, air and soil). To calculate it, three factors are needed: Effect Factor (EF), Fate Factor (FF) and the Exposure Factor (XF), as shown in Equation 1:

$$CF = EF \times FF \times XF \quad (1)$$

where FF accounts the persistence of chemicals in environmental compartments. The larger the FF, the harder the chemical can be removed from environmental compartment. XF accounts the likelihood of exposure and accumulation of chemicals to species, for aquatic ecotoxicity for example, XF is usually representing the dissolved chemicals in water[8]. The EF measures the toxicity of chemicals, which are often expressed in the response of species to certain level of exposure of chemicals[9,10].

The Data Gap in LCA. Since LCA considers the impacts and resource used over the entire product life cycle, including raw material extraction, manufacturing, use and the end-of-life management, therefore, large amount of data is required for a LCA study. In the LCI for chemical productions, often hundreds to thousands of data points are needed, depending on the system boundary of the study[11]. In the stage of LCIA, each of these emission flows require a CF to be able to convert the mass of emissions to the environmental and human health impacts.

However, the problems are obvious. On one hand, the number of CFs in the current literature is limited. For instance, *USEtox*, developed by United Nations Environment Programme/Society of Environmental Toxicology and Chemistry (UNEP/SETAC), is one of the most prevalent impact assessment methods in the field of LCA, contains pre-calculated CFs for 3,077 organic and 27 inorganic chemicals[12]. Intergovernmental Panel on Climate Change (IPCC) published the Global Warming Potential (GWP) in different time horizons for few hundreds of chemicals[13]. On the other hand, new chemicals are emerging every day. Chemical Abstract Service (CAS) reports that over 144 millions of chemicals have been registered in their database as of 2018, and

thousands of new are being added everyday[14,15]. The data gap between the existing CFs and the large number of chemicals posed challenges to LCA practitioners. Beyond the pre-calculated CFs, estimation and proxy methods have been applied to fill in the data gap to complete the LCA studies.

Methods to Fill in Data Gaps in LCA. In general, two types of methods are often used to fill in data gaps in the field of Environmental Science, and in LCA studies. The first approach relies on deterministic or dynamics-based models and depends on our ability to write all the dynamical and physical processes in mathematical way, and to discretize them so that they can be solved numerically. People also called it "mathematical model". The second approach is empirical or data-based. It depends on the available data and how we choose to use it in a statistical way, so we can recognize a reasonable pattern from the data and make prediction. In LCA, because of the complexity of the system usually has, most of the methods fall in to the realm of the data-based model.

Previous LCA studies applied data-based predictive model to fill-in data gap. Marengo et al., used linear regression to estimate the carbon emissions in cement productions, and achieved satisfying performance[16]. Park et al., approximated the life cycle assessment of product concept with multiple regression analysis[17]. Pascual-González et al. used a combination of multi-linear regression and mixed-integer linear programming to predict the life cycle impacts in different environmental categories for chemicals[18]. Many other predictive models and proxy methods are also developed for LCAs[19–22]. These simple predictive models are easy to use and to be understood.

However, their accuracy and performance are often not satisfying, when the analysis targets become more complex, for example, fine chemicals[23].

Machine Learning, and the Remaining Problem. The development of computational techniques and machine learning made new opportunities possible for LCA researchers to develop better predictive models. Machine learning also belongs to the realm of data-based predictive models. The object of machine learning is to use computational methods to let the computer extracting meaningful pattern from large amount of dataset (training data). It is not a new concept. In 1980's the then-AT&T Bell lab already use Artificial Neural Networks (ANN), one of the machine learning models, to detect zip code on envelop. One of the first applications of machine learning in Environmental Science is in Meteorology. Glahn and Lowry compared past meteorological model predictions to corresponding records of the observed actual conditions to tune the model or adjust its forecasts. This is what so-called Model Output Statistic (MOS)[24]. Nowadays, machine learning has been successfully applied in many areas in Environmental Science. These applications can be classified in three subcategories depending on the type of data sources.

Sensor-data-based application, such as in hydrology, using measured rainfall data to predict the amount of streamflow. Maier and Dandy reviewed 43 papers applying ANN methods to hydrological problems[25]. Hsu *et al* applied Multilayer Perceptron Neural Network (MLP NN) to model the rainfall-runoff relation in the Leaf River Basin in Mississippi[26]. Walter *et al.* used ANN model simulating the observed annual mean surface air temperature variations during 1874-1993[27]. The predictors of their model were

equivalent CO_2 concentrations and tropospheric sulfate aerosol concentrations, as well as volcanism, solar activities and El Niño–Southern Oscillation (ENSO) index. As the results, the ANN model can explain 83% of the observed temperature variance, which is significantly higher than the regression analysis.

Experimental-data-based application, such as in toxicology, using experimental data of chemical toxicity and molecular structural information to predict the toxicity for new chemicals without going through experiment. This area is relevant to LCIA since the EFs are essentially the response of species to chemical exposure. Phillips *et al*. developed a system to suggest "candidate alternatives" for 41 functional uses, in which the chemical can be used to other functional areas and exhibit relative lower bioactivity. Their model was based on Random Forest (RF). And they evaluated structural and physico-chemical properties descriptors as the inputs.

Other applications, such as using machine learning in LCA. Wernet et al. used ANN model to predict the cumulative energy demand (CED), global warming potential (GWP) and eco-indictor 99 (EI99)[28]. They also compared the ANN performance with linear regression model and showed that ANN outcompeted liner regression model by up to 0.4 in R^2. Wernet *et al* also conducted follow studies and applied their ANN model to fill in data gaps in LCA studies for pharmaceuticals. They showed that the prediction results of machine learning model can be used as a proxy data when no better information available in LCA.

The studies above solved the data problems in many areas, including LCA, using machine learning model at certain extend. However, there are still challenges and unsolved problems.

(1) Lack of interpretability. One of the very common criticisms for machine learning models used in Environmental Science and LCA is lack of interpretability. Taking ANN as an example, which has been criticized for long by its "black-box" nature. The contributions and mechanism of each molecular descriptor to toxicity endpoints are vague. What's more, putting the problem of contribution aside, many descriptors used for model development are not interpretable to human at the first place. There are thousands of descriptors available for ecotoxicity purpose. Sometimes the descriptors involved in reported QSAR models are not clearly defined or identified. To overcome this, modelers should conduct feature selection when developing machine learning models.

(2) Lack of proper model validation. Almost every machine learning models are good at interpolation, but not doing very well at extrapolation. Therefore, the model performance report on one part of data might not reflects the true performance of the model. This problem become more serious as the amount of experimental data getting smaller. To overcome this, cross-validation should be conducted when adequate computational resource is allowed. Tropsha and Golbraikh recommended that the process of training and test set selection and external validation should be carried out a number of times to identify the ranges of external predictively of a model[29]. What's more, for better performance, training data should be well-distributed over the full range of endpoint values. However, many of the existing models are not following this practice.

(3) Lack of model applicable domain (AD). How to measure the model uncertainty is always one of the research focuses in machine learning. External validation, i.e., is sometime not enough given the limited amount of experimental values and the lack of diversity in chemical type. Only providing a single prediction results

without any uncertainty analysis will reduce the usefulness of the model. To overcome this drawback, model Applicable Domain (AD) should always be reported along with the model. AD has been defined as the "response and chemical structure space in which the model makes predictions with a given reliability"[30]. More similar of the testing data to the training data will decrease the uncertainty of the model, than less similar testing data to training data.

B. Intellectual Significance and Objectives

To facilitate LCA studies, and to overcome the problems in the existing proxy and predictive models, this dissertation seeks to develop advanced machine learning models and provide innovative methods for LCA practitioners to fill in data gaps from different perspectives and under various data scarce situations.

The second chapter in this dissertation seeks to answer the question that whether the advanced machine learning model can learn meaningful relationship between chemicals structure and the characterized life cycle assessment results. The output of this chapter will contribute six new predictive models that are developed in ANN, to estimate the characterized results of organic chemicals, in six impact categories: global warming (IPCC 2007), acidification (TRACI), human health (Impact2000+), ecosystem quality (Impact2000+), and eco-indicator 99 (I,I, total). This chapter uses the selected molecular descriptors as the predictors to estimate the characterized results. The fundamental feature selection methods, the model validation methods and the theory of model AD will be described in this chapter, which will setup the theory foundation for the following chapters. The outcomes of this chapter prove that machine learning model can be used to

predict the final characterized results for LCA directly, and appropriate AD measurement is important to understand the reliability of the results.

The third chapter in this dissertation seeks to answer the question that whether ANN can be used to predict the Species Sensitivity Distributions (SSDs) for organic chemicals using their chemical structure, therefore to calculate their EFs for LCIA. As described above, EF is one of the important parameters in LCIA. EF can be calculated from SSD. This chapter will contribute new predictive models in ANN to estimate the LC50 values for 8 aquatic species for organic chemicals. And new SSDs will be built from these predictive ecotoxicity data. The benefit of doing so is that more data can be used to train the ANNs since the experimental data for various species is abundant. Another innovation of this chapter is that the molecular descriptors will be selected through two-steps feature selection algorithm, and the contribution of each descriptors to ecotoxicity will be evaluated. This is the first attempt to do so in the predictive models for LCA. At the end of this chapter, to demonstrate the models developed in this chapter, the chemicals in ToX21 database are used as candidate chemicals to estimate the ecotoxicological impacts. The outcome of this chapter shows that machine learning models can be used to predict the intermediate values in LCIA.

The fourth chapter of this dissertation turns to the FF in LCIA. FFs are usually calculated by mathematical models. For example, *USEtox* takes several chemical properties as inputs. And existing proxy methods have already been provided in *USEtox*. This chapter seeks to answer the questions that whether replacing the default proxy methods in *USEtox* by advanced machine learning model can improve the uncertainty of the FF. The sensitivity of the *USEtox* model to the inputs will be analyzed. Machine

learning models for each chemical property will be developed. The default or new machine learning methods that exhibit the narrowest uncertainty ranges will be used as the "best practice methods" to estimate the missing chemical properties, and then to calculate the FFs. Since the EFs can be estimated by the SSD models in the third chapter in this dissertation, new characterization factors for organic chemicals can be predicted by combining them together. This chapter will contribute 383 CFs for organic chemicals predicted by the model developed in this dissertation that are new to the literature. These CFs are reliable as they fall inside of the model ADs. This chapter is also the first attempt to understand the uncertainty of the CFs calculated by *USEtox*.

In conclusion, the increasing number of new chemicals to be evaluated by LCA brings up the need of advance estimation techniques to fill in the data gap in timely and accurate manner. Machine learning models, which have shown successes in many areas, provide new venture for LCA practitioners to tackle this challenge. My PhD dissertation seeks the linkage between machine learning and LCA. Together, the three chapters in my study examined the linkages from three different perspectives. With the advance machine learning models provided in this dissertation, LCA studies can be conducted at screening level when data is limited. The feature selection algorithm, and the model applicable domain analysis provide innovative ways to develop trustful models, and to validate the model performances.

II. Rapid life-cycle impact screening for decision-support using artificial neural networks

Abstract. The number of chemicals in the market is rapidly increasing, while our understanding of the life-cycle impacts of these chemicals lags considerably. To address this, I developed deep Artificial Neural Network (ANN) models to estimate approximate life-cycle impacts of chemicals. Using molecular structure information, I trained multilayer ANNs for life-cycle impacts of chemicals using six impact categories, including cumulative energy demand, global warming (IPCC 2007), acidification (TRACI), human health (Impact2000+), ecosystem quality (Impact2000+), and eco-indicator 99 (I,I, total). The Application Domain (AD) of the model was estimated for each impact category, within which the model exhibits higher reliability. I also tested three approaches for selecting molecular descriptors and identified the Principal Component Analysis (PCA) as the best approach. The predictions for acidification, human health and the eco-indicator 99 model showed relatively higher performance with R^2 of 0.73, 0.71 and 0.87, respectively, while the global warming model had a lower R^2 of 0.48. This study indicates that ANN models can serve as an initial screening tool for estimating life-cycle impacts of chemicals for certain impact categories in the absence of more reliable information. Our analysis also highlights the importance of understanding ADs for interpreting the ANN results.

A. Introduction

Chemical regulations increasingly focus on the product life-cycle aspects rather than end-of-pipe of production facilities. The Safer Consumer Product (SCP) program in

California, for example, requires that for priority chemicals under certain applications manufacturers must conduct alternative assessment taking into consideration the likely life-cycle impacts of the chemicals[31]. As a result, life-cycle assessment (LCA) is increasingly recognized as one of the tools for assessing alternatives in chemical design[32–34].

However, the pace at which LCAs are conducted cannot keep up with the pace at which new chemicals are developed. According to the Chemical Abstracts Service (CAS), there were over 100 million unique substances registered since June 2015, and about 15,000 new chemicals are added to the list every day[5]. The candidate chemical list of SCP alone contains over a thousand chemicals, each of which may require a full LCA study if there is growing concern about its use in consumer products[35]. The details of new and emerging chemical synthesis are considered highly protected intellectual property that is rarely disclosed to LCA practitioners, further limiting our understanding of their impacts[36].

Streamlined LCA approaches have been developed and tested to overcome this challenge[37–40]. Such approaches help screen the life-cycle impacts of chemicals without requiring extensive data[41]. Among others, the use of proxy data and regression models are two of the most common approaches to address the lack of data in LCA[17,18,42,43]. For example, proxy data were used to fill in the data gaps on bio-based products,[42] and linear regression models were used to approximate the carbon dioxide emissions from power plants[43]. These methods provide a way to fill in the data gaps at varying levels of uncertainty[13, 17, 18].

Another approach to the data gap challenge is the use of machine learning techniques, where molecular-structure-models (MSMs) are used to approximate the environmental impacts of chemicals. MSMs are widely applied in the Quantitative Structure–Activity Relationship (QSAR) field, where the chemical toxicity and physicochemical properties are estimated based on the chemicals' molecular structures[46–48]. The presence of inherent relationships between molecular structures and potential life cycle impacts of chemical enables MSMs estimate chemical life-cycle impacts using molecular structure information[23]. For example, chemicals with long chains, such as polymer, usually require multiple synthesis steps to bond small molecules together and requiring more energy, which in turn are more likely to generate CO_2 emissions and increase global warming impacts throughout the life cycle[49]. Similarly, the presence of nitrogen in the chemicals such as polyurethane indicates the use of nitrogen as an input, which increases the likelihood of nutrient emissions, increasing the potential of eutrophication impact[50]. Although in some cases these relationships are not intuitive or obvious to humans, a well-trained MSMs demonstrates its ability to estimate chemical life-cycle impacts[23].

Wernet and colleagues, for example, applied Artificial Neural Networks (ANN) with one hidden layer, which is one of the approaches in MSMs, to estimate the cumulative energy demand (CED) of pharmaceutical and petrochemical products[22, 24]. The authors also applied the technique to predict global warming potential (GWP), biochemical oxygen demand (BOD) and chemical oxygen demand (COD), with molecular structure descriptors as input to the models[52]. Comparing the model performance of ANN to that of linear regression, the authors showed that ANN with a

single hidden layer outperformed a linear regression model in estimating life-cycle impact indicators. However, the predictive power of these MSMs was still hindered by the lack of well-defined model training procedures, as well as the absence of uncertainty characterization of model outputs for new chemicals. Moreover, these ANNs can be further extended using multiple hidden layer.

In this study, I designed a novel approach for rapid screening of chemical life-cycle impacts based on ANN models and tested their performance. Our approach is the first effort to examine the application of ANN with multiple hidden layers in predictive LCA studies, and was developed using training, validation and testing techniques which are widely considered as the state-of-the-art in MSM[25, 26]. I determined differences in model performance when different sets of molecular descriptors were used as inputs to a given ANN model. Furthermore, I also characterized the confidence level of the ANN model outputs using the concept of Applicable Domain (AD), applied for the first time in the context of predictive LCA.

This paper is organized as follows: the 'Materials and Method' section presents the ANN model and the organization of the data used; the 'Results and Discussion' section shows the numerical results of the training, model application and the applicability domain, as well as interpreting the results; the limitations of the model, and future research directions are also discussed at the end of this paper.

B. Materials and Methods

Artificial Neural Networks. ANN is nonlinear, universal approximation models that usually have greater predictive power than linear regression and significant adaptability

for different tasks[55–57]. An ANN model consists of input, output and hidden layers. Within these layers are hidden neurons with activation functions, e.g., sigmoid or rectified linear unit (ReLU) function,[58] to project input data to nonlinear spaces. This allows ANN to solve problems that a simple linear regression model cannot. The layers are connected by weights that are trained during the training process. I then minimize the cost function, which measures the difference between predicted and observed values using the training dataset, by adjusting the weights. Therefore, the weights between layers will be updated during training to optimize the model prediction. An ANN model with more than one hidden layer is referred to as a deep Neural Network, which has recently become an important approach in the field of Artificial Intelligence (AI) and machine learning[59,60].

In our study, the input layer of the ANN model consists of molecular descriptors, which are numerical parameters with values that characterize various aspects of the chemical structure. The output layer generates a single characterized result for one impact category. The hidden layers serve to approximate the relationships between input and output layers. The final model is a system of fully interconnected neurons between a small number of hidden layers (one to three hidden layers), which is illustrated in Figure 1. This type of model structure is able to provide adequate predictive power with a shorter training time than more complex neural networks[61]. The ANN models in this study were developed using the Tensorflow framework in Python 2.7 under the Ubuntu 16.04 LTS system[62].

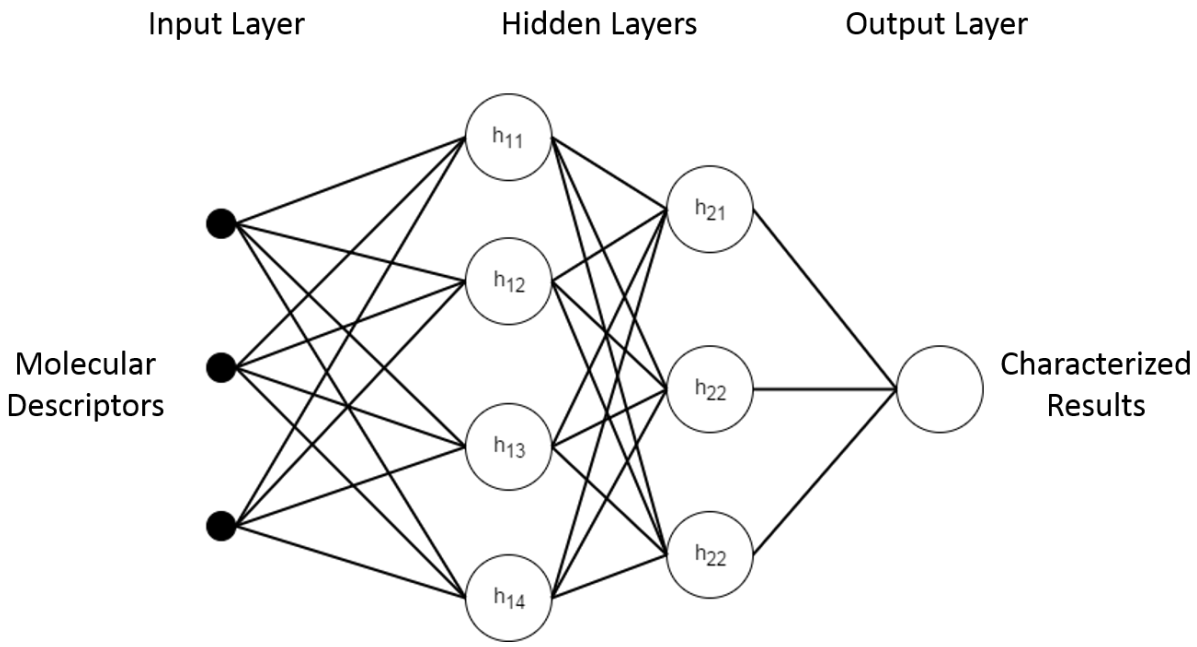

Figure 1. A conceptual diagram for a fully connected ANN model with two hidden layers. The solid lines between layers represent weights that are used in the approximation functions. The value in each node in the hidden and output layers is the sum of the values in the previous layer multiplied by the corresponding weights with appropriate activation functions.

Data Collection and Preprocessing. Training an ANN model is a supervised learning task, which means that both predictors and training targets must be included in the training process. In our study, I collected 166 unit process datasets for pure organic chemicals from the Ecoinvent v3.01 life-cycle inventory (LCI) database[63]. These chemicals were split into three groups for model development, optimization and reporting: training, validation and testing.

I selected three midpoint impact categories: cumulative energy demand (CED),[64] global warming (IPCC 2007, 100a),[65] acidification (TRACI 2.0);[66] and three endpoint

impact categories: eco-indicator 99 (I,I, total) (EI99),[67] ecosystem quality (Impact 2002+),[68] and human health (Impact 2002+).[68] The detailed explanations of these impact categories can be found in the supporting information. These six impact categories were chosen to test diverse aspects of a chemical's environmental impact.

Molecular descriptors are a critical component of the training data for our model. These descriptors are widely used in computational chemistry and the QSAR field to describe molecular structure.[69] Common descriptors are, for example, molecular weight, number of aromatic rings, number of functional groups and number of halogen atoms[70]. I used the software *Dragon 7* to calculate the molecular descriptors for the chemicals in this study[71]. *Dragon 7* calculates about 4,000 molecular descriptors for each chemical,[72] including constitutional, topological, ring and other descriptors. The large number of molecular descriptors generated by *Dragon 7* would make the training inefficient and could lead to the problem of overfitting[73]. It is therefore crucial to reduce the number of dimensions and extract an informative subset of descriptors. Several feature extraction and feature selection methods have been considered in the past[74]. Principal Component Analysis (PCA), for example, projects the descriptors to lower dimensions. PCA has been used in the context of developing predictive models using ANN[75–77]. The variables projected after PCA lose the physical meaning of the original molecular descriptors, but they do preserve most of the variance in the original dataset. Filter-based feature selection is another method, which removes descriptors with low variance and high mutual correlation. In the filter-based method, the remaining descriptors will preserve the physical meaning of the original descriptors; however, the removed descriptors might contain useful information for the predictive model. Therefore, filter-based feature

selection might affect the performance of the model. Another feature selection approach is the wrapper-based feature selection. This method conducts an extensive search to find the best subsets of molecular descriptors and selects the best subset according to the model performance. Due to its high computational cost and the risk of overfitting, I did not consider the wrapper-based feature selection method in this study[78].

In this study, I ran and compared the performance of three modeling cases: (1) using all descriptors generated by *Dragon 7* without any dimensional reduction; (2) using the descriptors selected by filter-based methods; and (3) using the features extracted by PCA that preserve 95% of the variance in the original dataset. The number of selected descriptors or features is the about same between the second and the third cases.

To achieve better model performance, each molecular descriptor selected after feature selection or PCA was normalized by calculating the z-score of each descriptor, as shown in Equation 2, to have zero mean and unit variance[79].

$$Z = \frac{X - \mu}{\sigma} \tag{2}$$

where Z is the descriptor after standardization, X is the original descriptor before standardization, μ is the mean value of the descriptor across all chemicals, and σ is the standard deviation of the descriptor across all chemicals.

Model Optimization and Validation. ANN models were trained for each of the six impact categories. Many hyper-parameters affect the performance of the final ANN model, such as the number of hidden layers, the number of hidden neurons in each hidden layer, and the learning rate during training[55]. Tuning each hyper-parameter is very time

consuming and, in many cases unnecessary. In our study, I optimized the number of hidden layers, as well as the number of hidden neurons in each hidden layer using the validation and test datasets. This ensured that the best model structure was used and that the model performance was not affected by the selection of the validation dataset[80].

To find the best hyper-parameters and model structure, ten chemicals out of the total 166 chemicals were randomly selected as the testing data, and 16 chemicals, or 10% of the remaining 156 chemicals, were used as validation data to report the model performance for training and optimization of the hyper-parameters in the ANN model. The other 146 chemicals were used as training data. The summary of the dataset used in this study is presented in Table S1.

Model Applicable Domain. Supervised-learning models make predictions based on what the models learn from the training data[61]. In general, models perform well on new chemicals that are structurally similar to the training data. Therefore, it is important to define the model AD so that the users understand the space within which a given model generates more reliable estimates.

Different AD measurement methods are available and discussed in the QSAR literature[30,81,82]. Based on the chemical LCI data collected in our study, I applied the Euclidean distance-based AD measurement method[81]. Other AD measurement methods, such as the probability density approaches, were not applicable to the data I collected in this study[30]. The Euclidean distance-based method measures the Euclidean distance in the descriptors' space from the query chemical to the mean of the training dataset, namely the training data centroid. This distance is defined as:

$$D = \sqrt{\sum (X_i - \mu_i)^2} \quad (3)$$

where D is the distance between the query chemical X and the training data centroid u; X_i and u_i are the i^{th} molecular descriptors of the query chemical and the centroid, respectively. Figure S3 illustrates the idea of distance-based AD measurement.

The confidence level of the estimation depends on whether the distance of the testing dataset to the centroid of the training data is smaller than a pre-calculated cut-off threshold. In many QSAR studies, this cut-off threshold is chosen subjectively by an expert judgement[30]. In our study, I selected the threshold in such a way that the difference between the average prediction error among the data points in the validation dataset within the AD and that among the data points outside is the largest. I then applied the selected cut-off threshold to the testing dataset.

C. Results

Chemical Used for Model Development. The chemical dataset I collected in this study represents a wide range of chemical types, including but not limited to petrochemicals, chlorine-based chemicals, and pharmaceuticals. The detailed list of chemicals used in this study can be found in the Supporting Information (Table S2). The mean, standard deviation, minimum, median and maximum values of the characterized results for the six impact categories are shown in Table 1, for the entire dataset (166 chemicals). The distribution of the characterized results is presented in Figure S2. For the impact categories of global warming, human health and ecosystem quality, more than 60% of the chemicals have characterized results smaller than the average characterized result in the corresponding impact category. This right-skewed distribution means that fewer

chemicals can be used to train these three models within the range of higher characterized results. To address this, I transformed the characterized results of global warming, human health and ecosystem quality models to log scale before training.

Table 1. Statistics of the characterized results for the six selected impact categories

	CED (MJ/kg)	acidification (moles of H^+ eq./kg)	global warming (kg CO_2 eq./kg)	EI99 (points/kg)	human health (DALY/kg)	ecosystem quality (PDF·m^2·year^{-1}/kg)
Mean	91.5	1.2	4.8	0.4	5.5×10^{-04}	9.8×10^{-05}
Standard Deviation	41.3	1.0	10.2	0.4	5.1×10^{-04}	9.6×10^{-05}
Minimum	19.9	0.1	0.0001	0.01	4.8×10^{-05}	1.3×10^{-06}
Median	85.2	1.0	3.2	0.3	4.3×10^{-04}	6.6×10^{-05}
Maximum	288.1	6.8	107.9	2.6	3.3×10^{-03}	4.9×10^{-04}

Comparison among the Approaches to Reduce the Dimension of Molecular Descriptors. Figure 2 shows the performance of the ANN model for predicting acidification, considering the validation dataset, based on: (1) all the descriptors generated by *Dragon 7* (3,839 descriptors), (2) descriptors selected with filter-based methods (58 descriptors) and (3) descriptors extracted by PCA that preserved 95% of the variance in the original descriptor sets (60 features). I examined each of the three cases with one, two, or three hidden layer(s), and 16, 64, 128 or 512 hidden neurons embedded in each layer. The performance scores were reported as the regression coefficient, R^2, for the validation dataset without the testing dataset.

As shown in Figure 2, the ANN models for acidification developed using all the descriptors exhibited the lowest R^2 values (green bars). Although the discrepancy is not significant, descriptors extracted using PCA resulted in a better performance in 8 out of

12 models than the descriptors selected using the filter-based method. The acidification model with two hidden layers and 128 hidden neurons embedded in each layer had the highest R^2 (0.75). In this acidification model, the R^2 was 0.33, 0.60 and 0.75 for the validation dataset considering the full, feature selection, and PCA descriptors, respectively. The same analysis for the ANN models of other impact categories can be found from Table S4 to Table S9. For the 72 different model settings (6 impact categories, 3 levels of hidden layers and 4 levels of hidden neurons) tested in this study, the ANN models developed using PCA descriptors performed better in general, with higher R^2 values for 49 ANN models using PCA (68%) than those developed using all or feature-selection descriptors. Furthermore, for every impact category, the PCA-based ANN models had the best performance (highest R^2) on the validation dataset. As a result, I employed PCA as the approach to reduce the dimensions in the input data and to improve the ANN's performance.

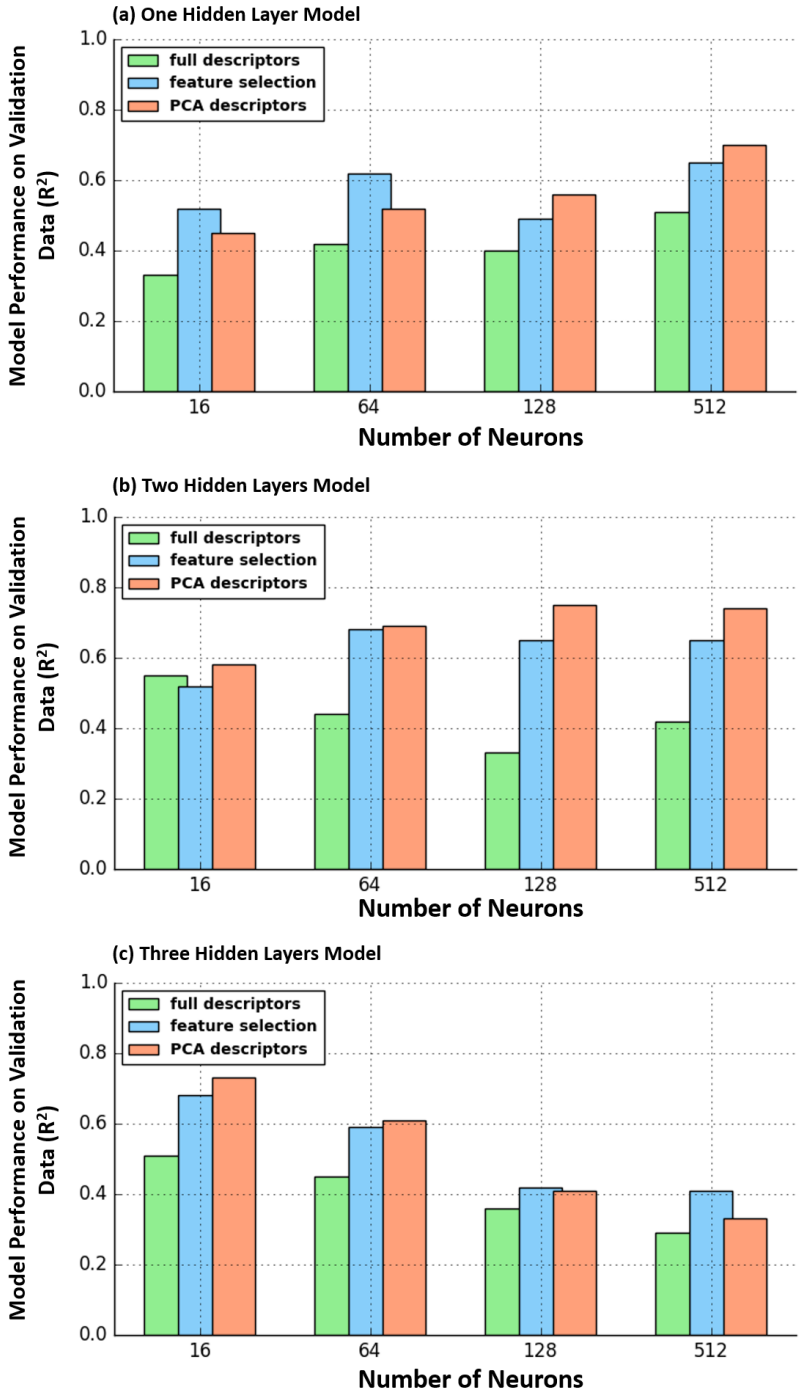

Figure 2. Performance (R^2) of the acidification model developed with: (1) all molecular descriptors set (green); (2) molecular descriptors after feature selection (blue); and (3) molecular descriptors after PCA (orange). The performances are the results using the

validation dataset without the testing dataset. The same analysis for the other models can be found from Table S4 to Table S9.

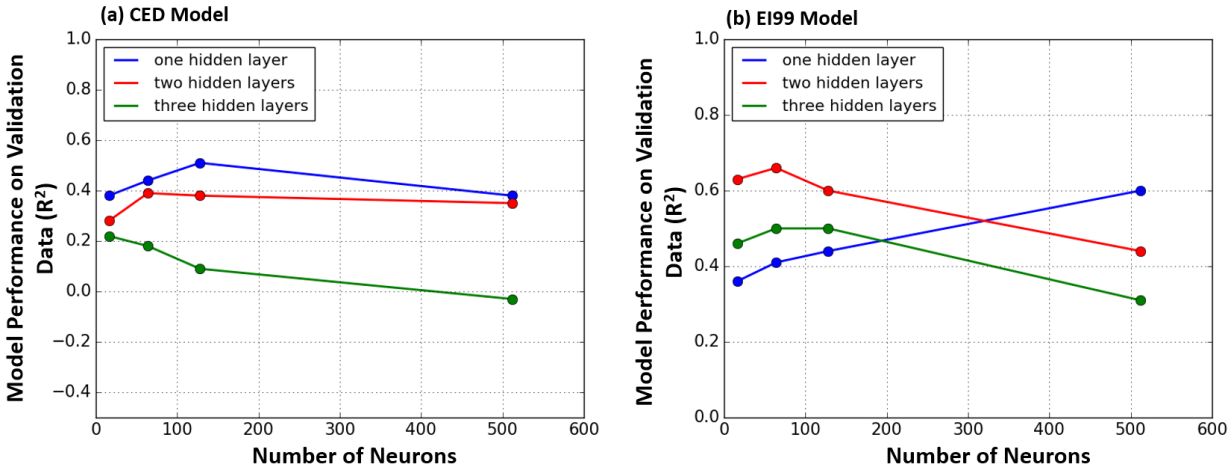

Figure 3. Model performance (R^2) using the validation dataset for (a): the CED model, and (b): the EI99 model with one, two and three hidden layer(s) and 16, 64, 128 and 512 hidden neurons embedded in each layer. Descriptors selected using PCA were considered as the input.

Figure 3 shows the results of optimization for the CED and EI99 models. The models were developed with the descriptors extracted by PCA and the performance was measured using the validation dataset. For CED, the model with one hidden layer and 128 hidden neurons in each layer showed the highest R^2 (0.51). For EI99, the model with two hidden layers and 64 hidden neurons in each layer showed the highest R^2 (0.66). Less complex models (e.g., the EI99 model with one hidden layer) did not have enough predictive power. However, due to the limited amount of training data, the model performance on the validation dataset decreased and overfitting occurred as I increased the complexity of the model. For both CED and EI99, the model with three hidden layers and 512 hidden neurons showed lower R^2 than less complex model settings (i.e., one or

two hidden layers). More training data will improve the model accuracy. However, inconsistencies and potential errors in the underlying LCI databases are limiting factors to the amount of training data I could collect.

Based on the validation results, the optimized model structure for each model is presented in Table 2. The human health model requires the highest complexity (three hidden layers with 64 hidden neurons in each layer) among all models. The details of the training process for each model, such as the learning rate, activation function and training epoch, can be found in Table S10.

Table 2. Optimized number of hidden layers and number of hidden neurons in each layer for the six models.

	Number of Hidden Layers	Number of Hidden Neurons in Each Layer
CED*	1	128
acidification	2	128
EI99**	2	64
global warming	2	16
human health	3	64
ecosystem quality	2	128

*cumulative energy demand;
**EI99: eco-indicator 99;

Model Performance. Six models were trained using PCA descriptors with the optimized model structure presents in Table 2 to estimate the characterized results for the six selected impact categories for organic chemicals. The performance of each model using the training, validation and testing datasets are reported (R^2 and Mean Relative Error (MRE)) in Figure 4 and Table 3. Each sub-graph in Figure 4 represents the model performance for the corresponding impact category. Circles represent the performance on the training dataset, the squares represent the performance on the validation dataset and

the triangles represent the model performance on the testing dataset. The solid diagonal in each graph represents the perfect prediction line, which is when the model prediction equals the reported value.

Among the six models, the acidification, EI99 and human health models perform relatively well, with R^2 of 0.73, 0.87 and 0.71 considering the testing dataset, respectively. The CED and ecosystem quality models showed lower performance, with R^2 of 0.45 and 0.48 on the testing dataset, respectively. The global warming model did not perform very well. Even though the R^2 on the testing dataset was 0.48, the training and validation accuracy were relatively low (0.31 and 0.21, respectively). This indicates that the global warming model still has room for further improvements.

Figure 4 also shows that chemicals with high life-cycle impacts tend to have higher estimation errors. This is because there is less training data available around such chemicals in the parameter space. In addition, chemicals with very high characterized results (especially for CED) are mostly pharmaceuticals (e.g., *pyrazole*). Their environmental impacts, such as energy intensity, are also affected by the selectivity and purity requirements of the pharmaceutical manufacturing process, in addition to their molecular structure. Therefore, their molecular structure is often insufficient to reliably predict the life-cycle impacts. This phenomenon would not be solved by simply increasing the model complexity. More training data from the pharmaceutical industry would be needed to solve this issue.

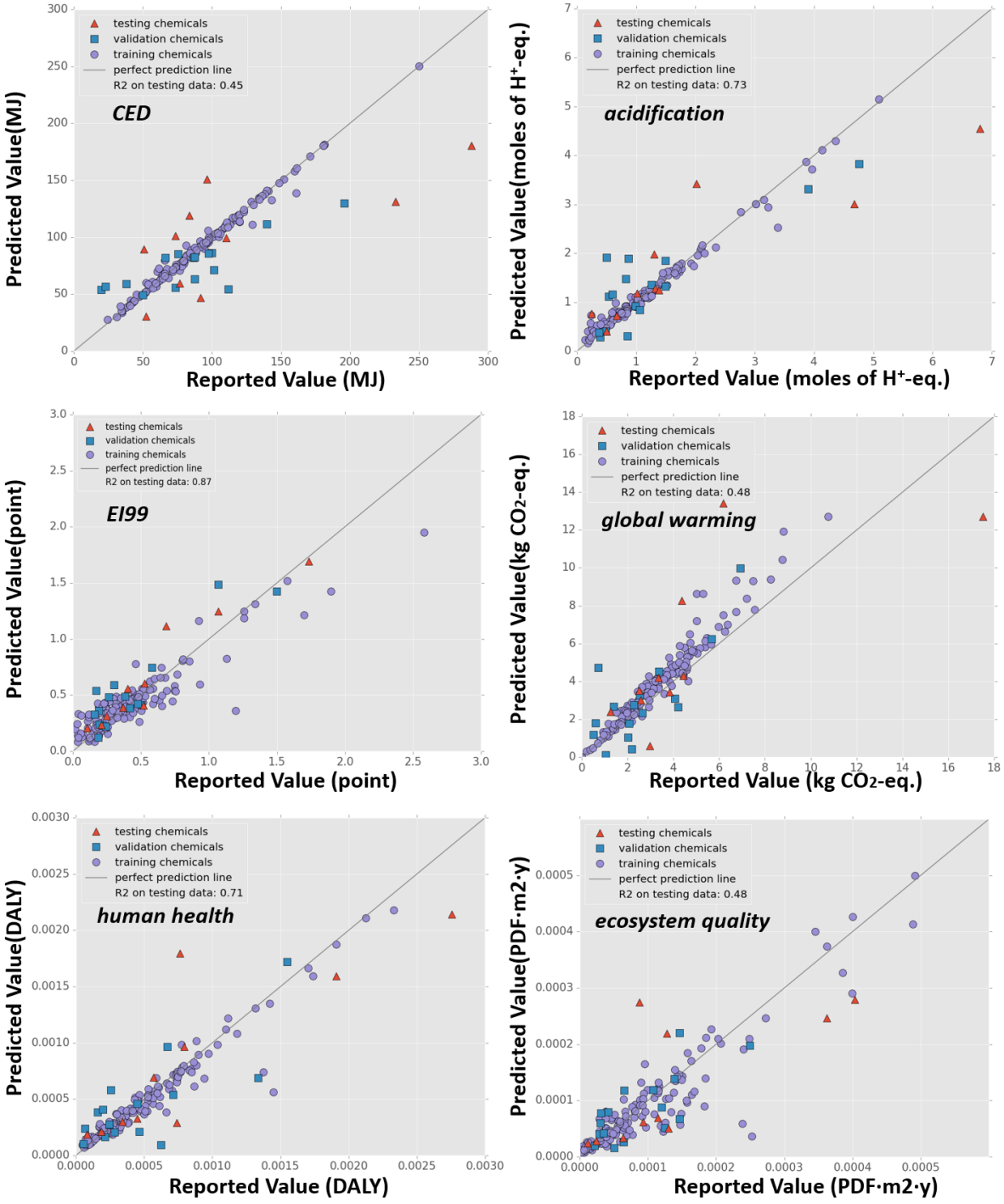

Figure 4. Model performance considering the training, validation and testing datasets. The training dataset was used to develop each model. The validation dataset was used to

optimize the model structure, and the testing dataset was used to report the model performance.

Table 3. Model performances for the training, validation and testing datasets

		CED*	acidification	EI99**	global warming	human health	ecosystem quality
Training Dataset	**R^2**	0.98	0.97	0.82	0.31	0.94	0.84
	MRE	3%	14%	55%	20%	15%	47%
Validation Dataset	**R^2**	0.52	0.75	0.72	0.21	0.58	0.48
	MRE	40%	56%	50%	88%	68%	52%
Testing Dataset	**R^2**	0.45	0.73	0.87	0.48	0.71	0.48
	MRE	40%	46%	30%	50%	46%	65%

*cumulative energy demand;
**EI99: eco-indicator 99;

Model Applicability Domain Analysis. The MRE of both the validation and testing datasets that fall within and outside of the AD in each model are presented in Table 4. The testing dataset within AD has a lower MRE than chemicals outside the AD for all models, except for global warming model. This shows that chemicals with higher Euclidean distance to the training data centroid tend to have higher prediction errors. Due to the limited performance of the global warming model, the predictions for chemicals with lower distance to the centroid also exhibit high errors.

Table 4. Mean Relative Error (MRE) of chemicals inside and outside of the measured AD on both validation and testing dataset for each model. The AD was measured on validation dataset.

	Validation Dataset		Testing Dataset	
	MRE within AD	**MRE outside AD**	**MRE within AD**	**MRE outside AD**
CED*	18%	47%	30%	44%
acidification	32%	150%	26%	76%
EI99**	36%	107%	21%	43%
global warming	25%	92%	65%	50%

human health	62%	180%	75%	111%
ecosystem quality	41%	104%	40%	63%

*cumulative energy demand;
**EI99: eco-indicator 99;

Case Study. I selected two chemicals, acetic anhydride and hexafluoroethane (HFE), from the testing dataset for a case study to demonstrate how our models. Acetic anhydride is an important regent for chemical synthesis, and HFE is an important industrial chemical for manufacturing semiconductors.

The estimation results for these two chemicals are reported in Table 5, along with the estimation error compared with the reported values, and the AD analysis results indicting if each chemical fall within the model AD. The AD of the global warming model was very narrow, and therefore both chemicals shown in Table 5 fell outside the AD. The reported values show that HFE has higher environmental impacts than acetic anhydride in all impact categories, and the model predictions successfully preserved this relationship, which is important when comparing the environmental impacts between the two chemicals. Overall, our models exhibited better performance for acetic anhydride than for HFE. The model with the highest error is the global warming model for HFE, with an absolute error of 116%. The estimation error for acetic anhydride is < 25% on the CED, acidification, global warming and EI99 models, while for HFE only the EI99 model has an estimation error lower than 25%. The AD measurement results successfully indicate that acetic anhydride falls within the AD for each model except for global warming model, and HFE is located outside of every model's AD.

Table 5. The model estimation results of acetic anhydride and HFE for the six selected impact categories in this study, along with the Applicable Domain (AD) analysis for these two chemicals. The numbers shows reported

values and the values in the parenthesis are values estimated by the model and the absolute value of relative error.

	acetic anhydride	hexafluoroethane
Within AD?	Yes*	No
CED (MJ)	83.8 (96.3, 15%)	232.9 (131.2, 44%)
acidification (moles of H^+ eq./kg)	1.0 (1.2, 16%)	6.8 (4.5, 34%)
EI99 (points)	0.4 (0.4, 6%)	1.7 (1.6, 6%)
global warming (kg CO_2-eq.)	3.3 (4.2, 25%)**	6.2 (13.4, 116%)
human health (DALY)	4.0×10^{-4} (5.2×10^{-4}, 30%)	2.7×10^{-3} (1.7×10^{-3}, 37%)
ecosystem quality (PDF·m^2·year)	9.3×10^{-5} (6.9×10^{-5}, 26%)	4.0×10^{-4} (2.6×10^{-4}, 33%)

* Excluding global warming model

** Out of AD

D. Discussion

The MSMs I presented in this study are not designed to be used for interpreting the mechanism between chemical structure and life-cycle impact. Instead, our model should be considered when there is a need to fill in data gaps or to screen life-cycle impacts of chemicals. The deep ANN models are known as "black-box" models, in which the contribution of each input variable to the final output values are not interpretable due to the large number of hidden neurons and multiple hidden layers embedded. Simple linear regression can be used to understand the mechanism and analyze the contribution of each molecular descriptor, but the prediction accuracy is much lower according to a previous study[28].

Since I use the existing LCI as the training data to develop the MSMs, the model estimations should be subject to all the assumptions and the uncertainties in the existing databases. It is well known that many chemical LCI datasets are derived using crude assumptions, heuristic rules, and stoichiometric relationships. The outputs of the models

using such data as the training dataset would provide comparable results with the existing datasets, since they cannot overcome the limitations of the datasets.

In our study, the Euclidean distance-based AD measurement was used to characterize the estimation uncertainty. Although this measure is shown to provide a reasonable indication of prediction errors, additional research is needed to derive uncertainty information using AD measures comparable to current LCA practice. Given the importance of the AD measures, the model confidence or uncertainty information should be more widely characterized and disclosed in predictive LCA research. Other model AD measurement methods, such as the non-parametric probability density distribution method, can be considered as a means to improve the AD measurement when training data is normally distributed[30].

Future research may be consider the synthesis pathway descriptors, such as reaction temperature, existence of catalyst or reaction selectivity as the model predictors instead of just using molecular descriptors. This will make the model more useful from the chemical engineering perspective. ANN can also be extended to the estimation of chemical LCIs in addition to characterized impacts, in which case LCA practitioners can use the characterization methods of their choice. Most of all, improving the availability of reliable and harmonized LCI data would be crucial to develop reliable ANN models for LCA.

E. Acknowledgements

The author acknowledges Prof. Arturo A. Keller and Prof. Sangwon Suh for his helpful comments over the development of the chapter. This publication was developed under

Assistance Agreement No. 83557901 awarded by the U.S. Environmental Protection Agency to the University of California, Santa Barbara. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication.

III. Expanding the coverage of species sensitivity distributions through artificial neural networks

Abstract. Species Sensitivity Distribution (SSD) is a key metric for understanding the potential ecotoxicological impacts of chemicals. However, SSDs were estimated for only handful of chemicals due to the scarcity in experimental toxicity data. Here we present a novel approach to expand the chemical coverage of SSDs using Artificial Neural Network (ANN). We collected over 2,000 experimental toxicity data in Lethal Concentration (LC50) points for 8 aquatic species, and trained an ANN model for each of the 8 aquatic species using molecular structure. The R^2 values of resulting ANN models ranged from 0.54 to 0.75 (median $R^2 = 0.69$). We applied the predicted LC50 values to fit SSD curves using bootstrapping method, generating SSDs for all 8,424 chemicals included in the *ToX21* database. We are making the code and the resulting SSD database open to the public. The dataset is expected to serve as a screening-level reference for understanding potential ecotoxicological impacts of chemicals

A. Introduction

Climate change, habitat losses and the exposure to various man-made chemicals are major threats to global biodiversity[83–85]. According to the Red List of Threatened Species by the International Union for Conservation of Nature (IUCN), 1,256 out of the total

8,455 threats are associated with pollution, of which 251 are due solely to the pesticide and herbicide[86].

Our understanding of chemical's toxicity footprints on the ecosystem, however, is limited by the sheer diversity of the chemicals used by the society, their wide variation in sensitivity across species, and the lack of experimental toxicity data[87,88]. The Chemical Data Reporting (CDR) of 2016 concluded that a total of 8,707 unique chemicals are produced or used in the U.S. in excess of about 11 tonne per year (for some chemicals a lower threshold was used) [89]. In 2018, the number of unique chemicals reported to have been produced or used in the European Union (EU) countries at the rate of one tonne per year or more reached 15,000 and growing[90]. Different species may exhibit dramatically different sensitivity to the same chemical; Pyrethroid, for example, is extremely toxic to insects, but it is well tolerated by most mammals[91]. An approach to estimate the potential ecosystem impacts of a chemical considering the variation in sensitivity of species to toxicants is the Species Sensitivity Distribution (SSD). SSD is a statistical distribution of toxicity data points (Lethal Concentration, or LC50, for example) across multiple species as a proxy measure for the ecotoxicological impact of a stressor to the entire community[92,93]. SSDs, combined with an assessment factor, are often used in risk assessment to estimate the Predicted No Effect Concentration (PNEC), which is usually the concentration at which five percent of the species are negatively affected (Hazard Concentration of five percent, or HC5)[94,95]. In environmental risk assessment, PNEC is often regarded as the safe concentration for chemical under which the entire aquatic ecosystem is unlikely be adversely affected[96,97]. Furthermore, SSD can be also used in Life Cycle Assessment (LCA), as the Hazard Concentration at which half of the species

are adversely affected, or HC50 value, is often used to derive the ecotoxicity Characterization Factors (CFs) of chemicals in Life Cycle Impact Assessment[12,98].

The challenge is that experimental toxicity testing data are scarce, while developing an SSD of a chemical requires multiple toxicity data points across multiple species[99]. The recommended minimum sample size ranges from 8 to 15[100,101]. The ECOTOX database, one of largest databases for experimental toxicity values, contains about 500 organic chemicals with experimental toxicity data for aquatic species, and only about 80 aquatic species have been tested on more than 5 organic chemicals In USETOX, which is one of the major models for chemical LCA, only about 2,000 experimental-based CFs exist for organic chemicals[12]. The scarcity of experimental toxicity data is the primary barrier for developing SSDs and for understanding the ecotoxicological impact of chemicals[102].

Quantitative Structure–Activity Relationship (QSAR) models have been used to approximate the relationship between chemical structure and their bioactivity or toxicity in the absence of available experimental data[103]. In the past decades, QSARs are often developed with simple models include liner regression or logistic regression for few species[104,105]. Mayer et al. for example, predicted chronic lethality of chemicals to multiple fishes using linear regression model from acute toxicity test data[106]. Raevsky et al. estimated the LC50 values of chemicals to *Guppy*, *Fathead Minnow* and *Rainbow Trout* using chemical similarity approach[107]. These QSARs, however, can only be applied on limited groups of chemicals, and failed to provide reliable prediction when applied on the others[103].

The development of machine learning techniques in recent years, however, opens an entirely new avenue of opportunities for developing predictive models in the fields where experimental data are scarce[108]. Artificial Neural Network (ANN), for example, has been successfully applied to predict rate constants and reaction rate of chemicals in atmosphere[19] and extreme weather,[109] and QSARs using simpler neural networks have also been used to estimate acute toxicity of chemicals to few aquatic species using inputs in varies formats. For example, Devillers developed QSAR model to estimate the acute toxicity of pesticide for *Lepomis macrochirus*[110]. Martin et al. provided a new model in Neural Networks to estimate the LC50 (96 hours) for *Fathead Minnow*, and achieved satisfying performance[111]. However, because of the development of SSDs require the ecotoxicity data in homogenous experimental condition and being tested on varies species taxa, the current studies failed to provide a group of homogenous models that can be used together to predict ecotoxicity data for multiple aquatic species in different taxa at once. Therefore, the existing QSARs from different studies can't be used together to generate trustful SSDs. What's more, many of the previous studies provided QSARs taking varies formats of model inputs, and some of inputs are difficult to be reproduced without extensive knowledge in corresponding areas.[110,110,112,113]

In this study, we present a novel approach to develop SSDs for organic chemicals using machine learning methods, taking only molecular structure as inputs. Experimental ecotoxicity data in LC50 for organic chemicals were collected for 8 aquatic species. ANNs using these data and their molecular descriptors were developed to estimate the ecotoxicity values in LC50, and therefore to build the SSDs for organic chemicals. A total of 8 ANN models were trained on experimental toxicity data for each of 8 aquatic

species: *Pimephales Promelas, Daphnia Magna, Oryzias Latipes, Oncorhynchus Mykiss, Lepomis Macrochirus, Cyprinodon Variegatus, Americamysis Bahia* and other water fleas. The performance of the predictive SSDs were evaluated on existing SSDs built by experimental data. The uncertainties of the ANN models as well as the predictive SSDs were analyzed. In the end, we applied our model and estimated the SSDs for over 8,000 organic chemicals in the Toxicology Testing in the 21st Century (*ToX21*) database and characterized their SSDs as well as the HC5 values. The performances of log-normal, Gamma and Weibull distributions to fit SSD were also evaluated.

B. Materials and Methods

Ecotoxicity Dataset Collection. We collected 2,521 experimental ecotoxicity data for non-ionizable organic chemicals on 8 aquatic species: *Pimephales Promelas, Daphnia Magna, Oryzias Latipes, Oncorhynchus Mykiss, Lepomis Macrochirus, Cyprinodon Variegatus, Americamysis Bahia* and other water fleas were collected from major public databases, including ECOTOX, eChem, EFSA and HSDB[114–118]. Data from peer-reviewed literatures was also added as supplementary data source to develop the neural network models in this study[107,110,111,119–122]. The number of experimental data collected for each species can be found in Figure S1 in supplementary information. To ensure data quality of the ecotoxicity dataset we collected from this study, the critical experimental conditions, such as the testing duration, chemical purity and *pH* values were strictly controlled during. 96 hours LC50 data was used for all species except water fleas (48 hours' data was used). Chemical purity must be higher than 85%. And the *pH* value must be in the range of 5 to 9. Experimental data that not meet these requirements was

discarded. For chemical with multiple experimental values, the geometric mean was used in the final dataset. To utilize some of the discarded data, and to increase the diversity of the species taxa, experimental values that met our data selection procedure for other water fleas in ECOTOX database was combined and treated as an individual species in this study. Within this category, there are 20 chemicals for *Ceriodaphnia Dubia,* 13 chemicals for *Daphnia Pulex and* 63 chemicals for *Mix Water Flea* (not specified). Additional information, such as the CAS number, SMILEs, molecular weight and the chemical names were also collected, for referencing purpose. The unit of the LC50 values were converted to *log10(LC50)* in $\mu mol/L$. The final dataset is available in the supplementary information.

Two-Steps Molecular Descriptor Selection. The original molecular structural descriptors were calculated using Python packages *rdkit* and *mordred*[123,124]. The descriptor calculators can produce over 2,000 descriptors for a single chemical, including basic physicochemical properties and autocorrelation descriptors. Large amount of descriptors could lead to overfitting problem[103,112]. Two steps feature selection procedures: filter-based plus tree-based feature selection, were used in this study to extract more meaningful descriptors.

Filter-based feature selection removes descriptors that have low variance, as well as the descriptors have high mutual correlations with others[66]. Tree-based feature selection method ranks the importance of each descriptor by their contribution to the prediction results in a decision tree model.[68] In this study, during the filter-based feature selection, descriptors with variance lower than 10 were discarded. Then, the correlations

between every leftover descriptor were calculated and the second descriptor was discarded if a descriptor pair has correlation higher than 0.6. A decision tree regressor in Python package *Sklearn* was used as the basis for the tree-based feature selection on the remaining descriptors[127]. The descriptors that contribute to the toxicity endpoint 3 times higher than the mean contribution were selected as the final descriptors in this study. As a result, The final descriptors are same for every chemicals for one species, but are different between species (different ANN models). In this study, we used 8 to 15 structural descriptors for developing our models. The most frequently utilized molecular descriptor was *SLogP* (Wildman-Crippen LogP), which appeared in all models. *Xp-2dv* (2-ordered Chi path weighted by valence electrons) and *PEOE_VSA6* (MOE Charge VSA Descriptor 6) were used in more than 3 models. The full list of descriptors used to develop each model in can be found in Table S6 of the supplementary information.

The Development of Neural Networks Models and Their Applicable Domain. ANNs were used as the modeling basis of the QSARs in this study. The ANNs were developed using *Tensorflow* and *Keras* in Python 2.7[128,129]. The hyper-parameters of ANNs that were optimized through five-fold cross-validation in this study, including the number of hidden layer(s), the number of hidden neuron(s) in each layer, the regularization factor and the type of activation function. These hyper-parameters were optimized by minimizing the mean square error (MSE) of the ANN models while holding others constant. The final models were built using the hyper-parameters that generated the lowest MSE during cross-validation. The final model performances were reported on 20

chemicals that were randomly selected and left out during model development for each species. The ANNs were built on the rest of data.

ANNs have better performance on inputs that are similar to the training data. We used Euclidean distance from the input descriptors to the centroid of our training data as the metric to evaluate the Applicable Domain (AD) in this study. The Euclidean distance is calculated as:

$$\boldsymbol{d_n = \sqrt{\sum(X_i - C_i)^2}} \quad (\mathbf{1})$$

where d_n is the distance of chemical *n* to the centroid of training data *C*; X_i and C_i are the i^{th} molecular descriptors of the input chemical and the training data. The centroid of the training data was calculated as the mean value of the molecular descriptors of all chemicals in the training data.

Whether an input chemical falls inside the model AD was determined by comparing a threshold value *K* with the distance d_n . For each ANNs, we first selected an initial *K* and then grouped the chemicals in the validation dataset by their distance to the centroid of the training data comparing with the *K* value (smaller or larger). The differences of the MSEs between these two groups were calculated. We then gradually increased the *K* value. The MSE differences changed accordingly since the chemicals within each group are different. We selected the *K* value that has the largest MSE difference to be the final threshold for model AD. The performance of this AD estimation was reported on the chemicals in the testing dataset.

The Development of SSDs and Their Uncertainties. SSD is a statistical distribution that illustrate the variation in the response of species to the exposure of chemicals. The

development of SSD begins with the generation of individual toxicity value of chemicals to species. In this study, we used LC50 values of chemicals to aquatic species. The LC50s are ranked from low to high, or the most sensitive to the least sensitive species. On the SSD graph, as shown on the Figure 2, the x-axis is the concentration of chemical, and the y-axis stands for the percentage of species affected. For each data point, the location on y-axis is the Median Rank position of it. Which is calculated using the *ppoint* function in *R*, and reproduced in *Python*[130].

Therefore, the LC50 values are used to estimate the Cumulative Distribution Function (CDF) of a selected distribution. Most of the SSDs were fitted using normal or log-normal distributions[131,132]. Other statistical method including log-logistic distribution and Burr Type III method are also exist but have not been widely used[132,133]. In this study, we used log-normal distribution as the basic distribution to fit SSDs, which was justified by the OVL analysis. The CDF of log-normal distribution is presented in Equation 4:

$$F_x(x) = \Phi\left(\frac{(\ln x) - \mu}{\sigma}\right) \tag{4}$$

where Φ is the CDF for a standard normal distribution $N(0, 1)$, shown in Equation 5, and μ and σ are the mean and standard deviation.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}x^2} \tag{5}$$

In this study, the decision of using log-normal distribution to fit SSD was made through running Overlapping Coefficient Analysis (OVL) testing on the screening results of *ToX21* database. OVL is a measurement for the similarity of distributions, which

compare the percentage of overlapping of the Probability Density Function (PDF)[134]. Equation 6 shows the mathematic representation of OVL for distributions $f_a(x)$ and $f_b(x)$:

$$\Delta\big(f_a(x), f_b(x)\big) = \int \min\{f_a(x), f_b(x)\}\, dx \tag{5}$$

For each chemical in the *ToX21* library, the actual distribution of the LC50 values on 8 species were compared with the empirical distributions that are fitted using the mean and standard deviation values on log-normal, Weibull and Gamma distributions. The area of overlapping was calculated.

Bootstrapping approach was used to estimate the uncertainty of SSD due to the limited amount of data points[135]. During each iteration of bootstrapping, eight data points were resampled using the fitted distribution curve and the newly sampled data points were used to construct new distribution curve. This process was repeated for 1,000 times, generating the upper and lower bounds of SSD for each chemical. The uncertainty of the QSAR predictions were also considered in the SSDs. Depending on whether the chemical fell inside or outside a model AD, different MSEs were attached to the QSAR predicted values. Therefore, the upper and lower bounds of SSDs can be reported.

Database Screening. The chemical list in the *ToX21* project is used as the candidates to be screened against the models developed in this study[136]. *ToX21* project aims to develop better toxicity assessment techniques in high-throughput robotic screening system. To date, 10,000 chemicals have been tested under the project, and the screening results help to identify chemicals for further investigation[136]. We removed inorganic chemicals,

ionized chemicals and chemicals that can't find SMILEs within this list. As a result, 8,424 chemicals are left and developed predictive SSDs using the models in this study. Among these chemicals, 1,239 chemicals fell into the ADs for more than 4 (out of 8) ANN models. We considered these predictive SSDs are trustful and discarded the rest of predictive SSDs.

HC5 values for these (1,239) chemicals were derived from the predictive SSDs. Among them, 218 chemicals were registered in the ECHA database, therefore we were able to find the production bands for them[137]. To consider ecotoxicity and production volume at the same time when comparing chemicals, we developed the concept of "Concern Index" in this study. The index is calculated as described in Equation 6. The screening results for all 8,424 chemicals, the "Concern Index" for the trustful 1,239 chemicals as well as their production band can be found in the supporting information.

$$CI = \frac{P}{HC5} \tag{6}$$

where CI stands for "Concern Index" (tonne·L/year·umol), which is a comparative score; P (tonne/year) is the annual production band reported in ECHA database; $HC5$ (umol/L) is the hazardous concentration read from the predictive SSD.

C. Results

ANN Model performances and Applicable Domain. The ANNs were developed using the optimized molecular descriptors, which were calculated and selected through feature selection algorithm before using them to train the model. After optimizing the number of layers and feature in our model, the performance of the ANNs ranged from 0.54 to 0.75 (mean 0.67, medium 0.69) in R^2 on the testing data. The performance of the ANN

model on *Americamysis Bahia* is presented in Figure 1 as an example. The performances of all 8 models, along with the number of hidden layers and neurons were summarized in Table 1. Other details about the model structure, including the activation functions and regularization factors during training can be found in the supplementary information.

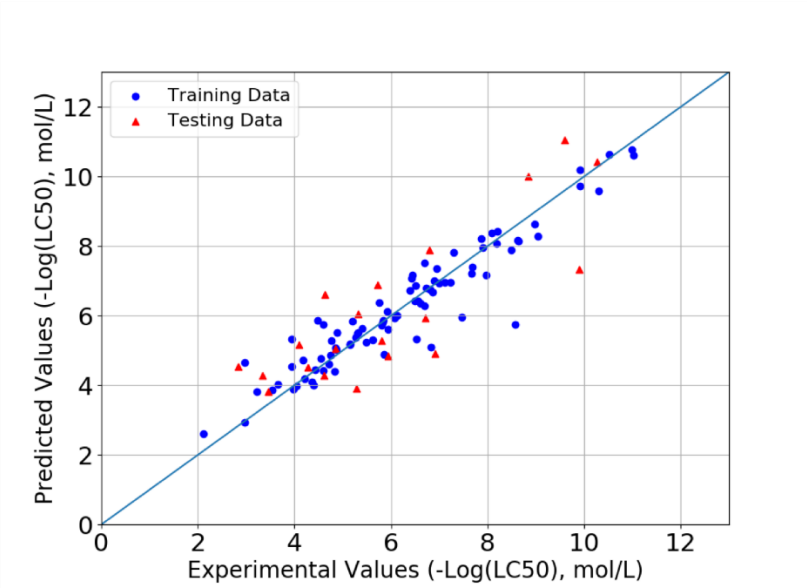

Figure 1. The performance of the *Americamysis Bahia* model on the training data (blue dots) and testing data (red triangles). The horizontal axis is the experimental values, and the vertical axis is the predicted values. The model structures were tuned using cross-validation technique. Information on other models is shown in the supplementary information.

The models for *Daphnia Magna* and *Oncorhynchus Mykiss* showed the highest R^2 on testing data (0.75), followed by the *Lepomis Macrochirus* (0.72) and *Pimephales Promelas* (0.71) models. The *Oryzias Latipes* model showed the lowest R^2 on the testing data (0.54).

Table 1. The performance of the ANN models on the testing data for the 8 aquatic species in R^2. The number of hidden layers and hidden neurons for each ANN model.

	*PP	*DM	*OL	*OM	*LM	*CV	*AB	*OWF

Model Performance (R^2) on Testing Data	0.71	0.75	0.54	0.75	0.72	0.66	0.67	0.63
Number of Hidden Layer	2	1	2	2	2	2	1	2
Number of Hidden Neuron in Each Layer	32 × 16	16	64 × 32	64 × 32	32 × 16	16 × 8	16	16 × 8

*Spcies acronyms: *Americamysis Bahia* (A.B.); *Daphnia Magna* (D.M.); *Lepomis Macrochirus* (L.M.); *Oncorhynchus Mykiss* (O.M.); *Cyprinodon Variegatus* (C.V.); *Oryzias Latipes* (O.L.); *Pimephales Promelas* (P.P.) and Other Water Fleas (O.W.F.).

We employed the concept of Applicable Domain (AD) to characterize the prediction accuracies of the ANN models and serves as a proxy to estimate whether a chemical is appropriate for the QSARs. The results of AD analysis are presented in the supplementary information. Among the ANN models that we developed, *Oncorhynchus Mykiss* and the *Lepomis Macrochirus* models have the narrowest ADs. For these two models, the mean square errors (MSEs) of the testing data inside of the ADs showed 6%, while those outside of AD were 15% and 22%, respectively. For the *Pimephales Promelas* model, however, the average MSEs inside and outside of AD were 8% to 220%, respectively, indicating limited utility of the model outside of AD.

Predictive Species Sensitivity Distributions and Evaluations. Using our ANN models, we were able to estimate LC50 values for 8,424 chemicals from the *ToX21* database for each of the 8 aquatic species. We also estimated the prediction errors of the ANN models, as well as the inherent error of SSDs due to the limited number of data points. These SSDs can be found in the supporting information. Given the large number of chemicals in our results, we randomly selected a few chemicals to compare our predictive SSDs with the SSDs derived from experimental data. Elaborated here is one of them, DCMU (*3-(3,4-dichlorophenyl)-1,1-dimethylurea*), an algaecide.

The predictive SSD for DCMU is shown in red line in Figure 2. The figure also shows the uncertainty range of the ANN-derived SSD in grey. This uncertainty range was calculated by the prediction error of each ANN model, which was determined by whether this chemical fell into the AD of each model or not. For a comparison, we collected experimental data for the same species, and we were able to locate experimental LC50 values for the same list of species other than *Oryzias Latipes*, which were unavailable in the literature and databases that we referred to. Using these experimental values, we constructed an SSD as shown by the green line in Figure 2. According to the SSD derived from experimental values, the HC5 of DCMU was about 1.82 μmol/L, whereas the HC5 from the ANN-based SSD ranged from 2.51 to 3.24 μmol/L. Both experimental SSD and the predictive SSD showed that *Pimephales Promelas* has the best tolerance to DCMU in water, with the experimental LC50 of 61.7 μmol/L and the predicted LC50 of 75.9 μmol/L.

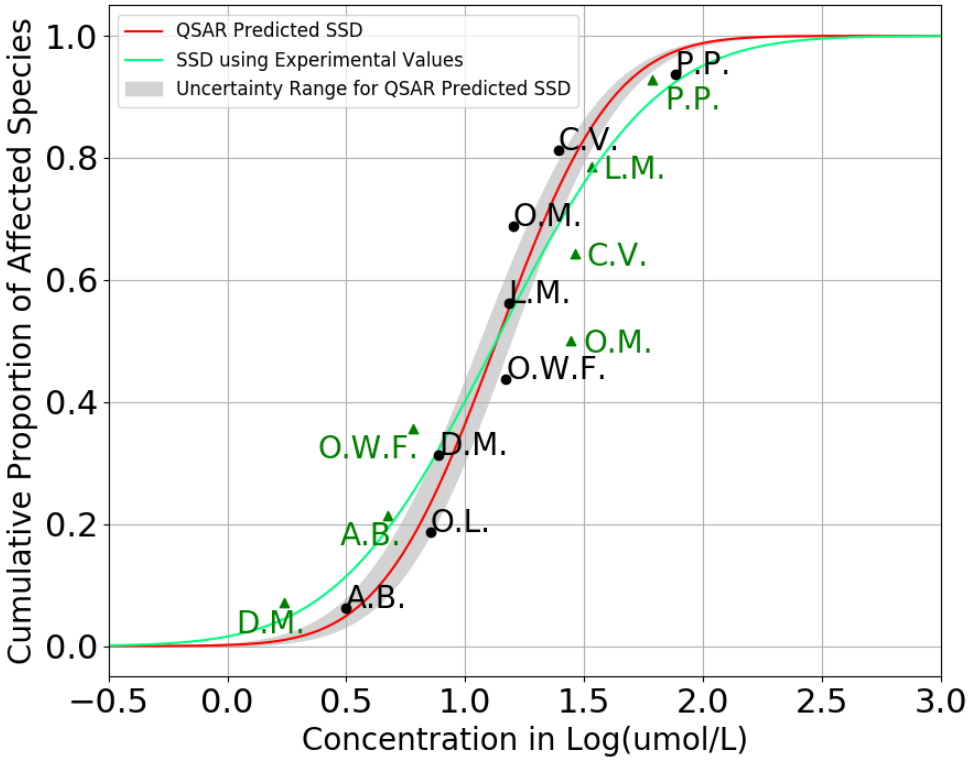

Figure 2. The SSD of DCMU (solid red line) constructed using the ANN-based LC50 values (black points), along with the uncertainty of ANN predictions (grey area) based on the model AD estimation for *Americamysis Bahia* (A.B.), *Daphnia Magna* (D.M.), *Lepomis Macrochirus* (L.M.), *Oncorhynchus Mykiss* (O.M.), *Cyprinodon Variegatus* (C.V.), *Oryzias Latipes* (O.L.), *Pimephales Promelas* (P.P.) and Other Water Fleas (O.W.F.). The SSD in green was constructed using experimental LC50 values found for 7 species).

Another 10 organic chemicals were randomly selected from the *ECOTOX* databases to evaluate the SSDs derived from our ANN models. We collected experimental LC50 data of these chemicals on other species than the aforementioned 8 species in order to avoid any overlap with the training data we used to develop our ANN models. Given the inherent uncertainty in SSDs due to the limited number of data points, we used the bootstrapping technique to visualize the potential range of SSDs. The mean, lower and upper bounds of HC50 (hazardous concentration for 50% of the species) values on both predictive and experimental SSD curves are presented in Table 2. The

Overlapping Coefficient (OVL) score in Table 2 shows the percentage of overlapping of the area of the predictive distribution and the experimental distribution. The detailed model prediction data for each of the chemicals, as well as the experimental LC50 values can be found in Table S4, and in the supplementary information. The predictive SSD, experimental SSD along with their overlapping area for chemical *chlorpyrifos* (*2921-88-2*) are presented in the supporting information as an example.

Table 2. The HC50 values of 10 chemicals in the ECOTOX database, along with the mean HC50 values for both ANN-based SSD and the experimental SSD, as well as the percentage of overlapping of the distributions based on the predictive and experimental SSDs.

Chemical CAS	Chemical Name	HC50 Mean (Lower, Upper Bound) in log(µmol/L)		OVL Score
		Predicted	Experimental	
50-29-3	*clofenotane*	-0.45 (-1.5, 0.62)	-0.85 (-1.43, -0.26)	70.6%
87-86-5	*pentachlorophenol*	0.32 (0.04, 0.62)	0.23 (-0.11, 0.54)	89.6%
58-89-9	*lindane*	1.29 (0.26, 2.22)	0.87 (0.36, 1.4)	65.8%
60207-90-1	*propiconazole*	0.64 (0.08, 1.25)	0.88 (0.5, 1.25)	75.9%
138261-41-3	*Imidacloprid*	2.1 (1.4, 2.8)	1.65 (0.66, 2.7)	77.6%
115-29-7	*endosulfan*	-0.46 (-1.09, 0.23)	-0.99 (-2.12, 0.1)	72.0%
2921-88-2	*chlorpyrifos*	-0.03 (-0.63, 0.66)	0.01 (-0.76, 0.84)	92.0%
206-44-0	*fluoranthene*	0.9 (0.22, 1.58)	0.23 (-0.04, 0.54)	50.3%
62-53-3	*aniline*	2.48 (2.21, 2.76)	2.71 (2.04, 3.42)	55.2%
333-41-5	*diazinon*	0.1 (-0.72, 0.91)	0.04 (-0.81, 0.87)	96.8%

Table 2 shows that the predicted HC50 values generated by the ANN models are generally in line with the experimental SSDs. The OVL results show that 8 out 10 chemicals have OVL score higher than 70%, which means that 70% of the area in the predictive SSD overlap with the SSD generated by the experimental data. Among them,

the predictive SSD for the chemical *diazinon* (*333-41-5*) share the largest overlapping area with the experimental SSD (96.8%), followed by the chemical *chlorpyrifos* (*2921-88-2*) by 92.0% overlapping area. The predictive SSD shows the lowest OVL score is the one for chemical *fluoranthene* (*206-44-0*) with OVL score 50.3%, and followed by the SSD for chemical *aniline* (62-53-3) with OVL score 55.2%.

We used the 97.5% percentile and the 2.5% percentile as the upper and lower bounds, respectively, of the 1,000 time bootstrapping when fitting LC50 values to SSDs. Mean values of predicted HC50 for all 10 chemicals were found within the upper and lower bounds of experimental counterparts, regardless of the species and number of data points. Figure 3 shows the mean SSD curves for chemical *chlorpyrifos* (*2921-88-2*), as well as the upper and lower bounds according to 1,000 times of bootstrapping (in light colors) for both experimental (red) and predictive (blue) SSDs. The range of experimental and predictive SSD are mostly overlapped according to Figure 3. The HC50 values of *chlorpyrifos* based on predictive SSD ranged from 0.23 to 4.57 µmol/L, and the experimental HC50 values ranged from 0.17 to 6.92 µmol/L. On both curves, fishes tend to be more sensitive to the exposure of *chlorpyrifos*. The species have the highest tolerance on the experimental SSD is *Sialis Lutaria* (Insects/Spiders) with LC50 61.66 umol/L, and on the predictive SSD is other water fleas with LC50 436.52 umol/L.

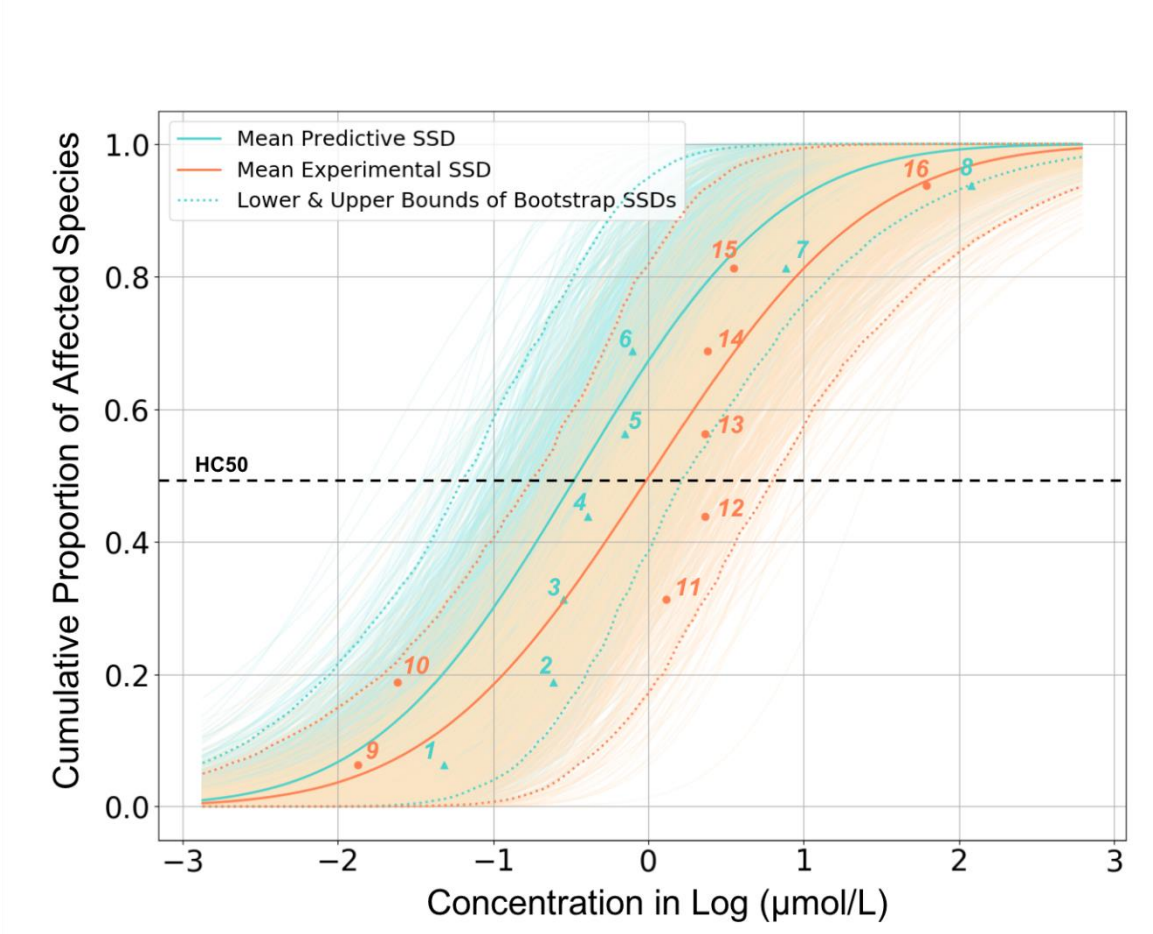

Figure 3. The mean (solid blue line), upper (97.5%) and lower (2.5%) bounds (dash blue lines) of the predictive SSD, and the mean (solid red line), upper (97.5%) and lower (2.5%) bounds (dash red lines) of the experimental SSD for *chlorpyrifos*. Each data point and numbers on the curves represents a species for corresponding data group (predictive, blue, or experimental, red). 1: *Americamysis Bahia* (Crustaceans, shrimp); 2: *Cyprinodon Cariegatus* (Fish); 3: *Daphnia Magna* (Crustaceans, water flea); 4: *Lepomis Macrochirus* (Fish); 5: *Pimephales Promelas* (Fish); 6: *Oncorhynchus Mykiss* (Fish); 7: *Oryzias Latipes* (Fish); 8: Other water fleas (Crustaceans, water flea); 9: *Pungitius Pungitius* (Fish); 10: *Gasterosteus Aculeatus* (Fish); 11: *Neocaridina Denticulate* (Crustaceans, shrimp); 12: *Lctalurus Punctatus* (Fish); 13: *Aplexa Hypnorum* (Molluscs); 14: *Carassius Auratus* (Fish); 15: *Zilchiopsis Collastinensis* (Crustaceans); 16: *Sialis Lutaria* (Insects/Spiders).

Screening the ToX21 Database. We applied the our models to the organic chemicals in the *ToX21* dataset to estimate the ecotoxicological impacts of these chemicals. As the

results, 8,424 organic chemicals are in the final dataset to be screened by our model. Among these chemicals, 1,240 of them fell into the AD for 4 or more models out of 8. Their predicted LC50 values, predictive HC5 and SSDs are provided in the supplementary information.

Using the screening toxicity results, we found the top 10 chemicals with the highest "Concern Index" in the registered chemicals in European Chemicals Agency (ECHA) database[137]. These top 10 chemicals are shown in Table 3. These chemicals are likely to raise concerns due to the high volume used and the high ecotoxicity (according to our screening results). More explanations about the methods we used in this screening analysis, as well as about the "Concern Index" can be found in the Methods section. The implications of these screening results are discussed in the Discussion section. The full screening results for the chemicals overlapped with the registered chemicals in the ECHA database can be found in the supplementary information.

Table 3. The top chemical chemicals with the highest "Concern Index" among the registered chemicals in the ECHA database.

Chemical Name	Chemical CAS	Concern Index (tonne·L/year·umol)	HC5 (umol/L)	Production Band in ECHA (tonnes/year)
4,4'-Diphenylmethane diisocyanate	101-68-8	504001.48	0.20	100000 - 1000000
2-Ethylhexyl acrylate	103-11-7	32449.21	3.08	100000 - 1000000
2-Ethylhexyl nitrate	27247-96-7	17868.23	5.60	100000 - 1000000
Anthraquinone	84-65-1	7988.09	0.13	1000 - 10000
tert-Butylperoxy 2-ethylhexyl carbonate	34443-12-4	5151.68	0.19	1000 - 10000
Dodecanoic acid	143-07-7	3878.86	2.58	10000 - 100000
2-Methyl-4'-(methylthio)-2-morpholinopropiophenone	71868-10-5	3411.48	0.29	1000 - 10000
Methyl dodecanoate	111-82-0	3029.26	3.30	10000 - 100000
6H-Dibenzo[c,e][1,2]oxaphosphinine 6-oxide	35948-25-5	2691.41	0.37	1000 - 10000

1,3-Benzenedicarboxylic acid	121-91-5	1751.49	57.09	100000 - 1000000

OVL Testing. SSDs can be fitted by different statistical distributions. We used the coefficient of overlapping (OVL) method to compare the performance of different statistical distributions: log-normal, Weibull and Gamma, when fitting SSD curves. As the results, the average OVL score of log-normal distribution was 0.817. More than 93% of the 8,424 SSDs have OVL score higher than 0.6 on log-normal distribution. The comparison between log-normal, Weibull and Gamma distributions is presented in supporting information. The average OVL scores for Weibull and Gamma distributions were 0.708 and 0.672, respectively. Log-normal distribution was the one that has the highest average OVL score among all distributions we tested. The resulting standard log-normal SSD function shows the average logmean (μ) and average GSD (geometric standard deviation, σ) of 3.21 and 2.58, respectively for the 8,424 SSDs.

D. Discussion

To our knowledge, our study is the first that consolidated aquatic ecotoxicity data from multiple data sources, and used them for large-scale SSD development using ANN. The resulting dataset, which is, to our best knowledge, the largest of its kind, is made freely available through our website. The predictive SSD, can be used in screening analysis to estimate the safety concentration of chemicals in aquatic ecosystem. Furthermore, LCA practitioners, who usually suffer from the absence of chemical ecotoxicity data[138], could

estimate the aquatic ecotoxicity for organic chemicals through the models developed in this study, therefore to calculate the Characterization Factors in impact assessment.

It is clear that our models cannot replace SSDs derived from experimental toxicity data. Given the current scarcity of experimental data and the high cost of developing them, however, we believe that our results demonstrated the potential for machine learning techniques to be used as a proxy for data gaps. Furthermore, the rapidly growing number of chemicals in the lab and in the marketplace makes it challenging for experimental data alone to meet the needs for understanding the potential ecotoxicological impact of chemicals. We believe that our results can serve as a pre-screening tool in the absence of experimental data to prioritize the candidates for further analysis. We view machine learning techniques not as a replacement of but as a complementary tool for experimental studies. We recommend that our results are used as a screening-level reference especially when experimental data is unavailable. High species sensitivity or low HC5 values in the our SSD database should constitute a reason for in-depth testing, while low species sensitivity or high HC5 values from our database alone should not be taken as a proof that the chemical is safe.

We demonstrated the screening ability of our model in the results of analyzing potential high ecotoxicity chemicals in the *ToX21* database, which also have high production volume according to ECHA database (Table 3). Among all chemicals, *4,4'-Diphenylmethane diisocyanate* (*101-68-8, MDI*) shows the highest "Concern Index", due to the ecotoxicity and the high production volume of it. *MDI* is widely used in the manufacture of *polyurethane*. *MDI* makes up about 60% in the global production of diisocyanate in 2000[139], and the U.S. demand for pure MDI was about 200 million

pounds in 2008[140]. MDI can be dangerous when used in consumer products and disposed inappropriately. *MDI* can be released from adhesive and sealants in a format that isn't completed reacted, therefore cause potential occupational exposure[140]. Record shows that MDI has the lowest ecotoxicity among *isocyanates*, but it can still cause side effects including skin irritation and respiratory failure[141].

We believe that the complementarity between predictive modeling and experimental studies can be further improved by standardizing the conditions for toxicity experiments and reporting. First of all, we cannot emphasize enough the importance of standard and machine readable data exchange protocol on experimental conditions. Due to the poor documentation and the lack of standard data exchange protocol, extracting data on experimental conditions from existing literature and databases required painstaking effort. Second, consistency in experimental conditions is crucial. We could not utilize many valuable experimental data points because one or more experimental conditions were not identical to the rest of the dataset. The variation in experimental conditions in e.g., duration of exposure, temperature, and chemical purity, significantly degraded the value of experimental toxicity data. A wider adoption of standard protocol for documenting and sharing toxicity testing results is urgently needed to tap into and maximize the value of experimental toxicity data for predictive modeling. While there are existing standards and guidelines including the OECD Test Guidelines, the Good Laboratory Practice (GLP) principles, and the Catalogue of Standard Toxicity Tests for Ecological Risk Assessment (REF), a universal applicable testing guideline is still lacking.

Machine learning techniques for ecotoxicological applications are still in a nascent stage, and there are large rooms for improvement on our study. Experimental data in better quality and quantality will improve the performances of the ANNs. Our models do not properly represent the toxicological impacts under multi-stressor conditions, because the experimental data used for training our model are all based on single chemical species. In fact, mixtures of chemicals are scarcely tested for ecotoxicity, and the development of protocols for mixture testing and reporting is in its infancy. In reality, however, ecosystem species are exposed to multiple chemicals at any given time. Although there are some researches confirmed the concentration addition effect of chemical mixture[142–145], given that the number of possible combinations of chemical mixtures in both composition and proportion is infinite, experimental data alone cannot be relied upon. Additional data and researches are needed to adequately address the ecotoxicological impacts of multiple stressors, especially in the context of using SSDs.

E. Acknowledgements

The author acknowledges Prof. Arturo A. Keller and Prof. Sangwon Suh and Prof. Dingsheng Li for his helpful comments over the development of this chapter. The author would also like to thank Alexander Chang and Mengya Tao for their help on collecting data, and Yuwei Qin for her help on the development of OVL. This publication was developed under Assistance Agreement No. 83557901 awarded by the U.S. Environmental Protection Agency to the University of California, Santa Barbara. It has not been formally reviewed by EPA. The views expressed in this document are solely

those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication.

IV. Reducing the Uncertainty of the Characterization Factors in USEtox by Machine Learning – A Case Study for Aquatic Ecotoxicity

Abstract. Life Cycle Impact Assessment requires the knowledge of chemical fate. *USEtox* contains a well-establish model to evaluate the Fate Factors (FFs) of chemicals, and several chemical properties are required. While the default proxy methods are provided by *USEtox* when the experimental data is in absence, the uncertainties introduced by these proxy methods are remain unknown. Here we present a study that aims to replace the default proxy methods in *USEtox's* fate model by machine learning models. New models in Artificial Neural Networks (ANNs) and Random Forest (RF) were developed for chemical properties including octanol/water partition coefficient (K_{ow}), biodegradation rate in water (k_{degW}) and others as the inputs to the fate model in *USEtox*. The errors of both default proxy methods and new machine learning models were assessed by comparing them with the experimental values, and the best practice methods to run the fate model were recommended. The uncertainty range of the *USEtox* FFs and Characterization Factors (CFs) were evaluated by Monte Carlo Simulation (MCS). The result shows that the standard deviation of the FFs using best practice methods ranges from 9.54 to 380.19 kg/kg·days^{-1}, while using default proxy methods ranges from 1.58 to 630.96 kg/kg·days^{-1}.

A. Introduction

USEtox is a well-established impact assessment model that aims to estimate the Characterization Factors (CFs), which are used to quantify the adverse environmental impacts caused by unit chemical emissions to different environmental compartments as a toxicity indicator in Life Cycle Assessment (LCA)[12,146]. The fate model is one of the components in USEtox that aims to evaluate the persistency of chemical emissions in compartment. The output of the fate model: the Fate Factor (FF) can be used together with Effect Factors (EF) and Exposure Factors (XF) to calculate the CFs for human and aquatic ecotxicity[12,147]. This fate model, along with the other models in USEtox, represents the best scientific consensus of the Life Cycle Initiatives since 2002. Several literatures have been published describing the methodologies of the fate model[148,149].

Several chemical properties are needed to calculate the fate model in USEtox, including octanol/water partition coefficient (K_{ow}), organic carbon-water partitioning coefficient (K_{oc}), biodegradation rate in water (k_{degW}), vapor pressure under 25 ℃ (Vap25), solubility in water under 25 ℃ (Sol25) and others[12]. In the latest version of USEtox, CFs for 3,077 organic chemicals were provided, in which the FFs were calculated using either experimental values, or scrutinized proxy value. When the FFs for new chemicals need to be calculated while there aren't experimental values for them, USEtox asks the users to refer to the proxy methods provided by EPIsuite, which is a tool developed and assembled by US EPA to predict serval chemical property endpoints[150]. Although those default proxy methods are widely used and well-established for many years, whether their accuracies can be improved, and what is the uncertainty of the USEtox FFs using those predicted chemical properties are remain unclear to us.

The default proxy methods were developed using the relationship between different chemical properties as well as the "fragment constant" methods[151,152]. Other approaches based on machine learning to estimate chemical properties have been already conducted in literature. Allision, for example, used Neural Network based model to estimate the OH rates in atmosphere and therefore predicted the Global Warming Potential (GWP), and reduced the uncertainty compared with other estimation methods[19]. Shafiei et al., used machine learning approach to estimate the solubility of hydrogen sulfide in ionic liquids, and showed promising accuracy in the process gas sweetening[19]. Cheng et al., developed an additive model, which requires some knowledge from the user about the target chemical, to estimate K_{ow} for organic chemical and showed good accuracies[153]. These studies, along with the others, showed promising outcomes when using machine learning based methodologies to estimate chemical properties for the application in chemistry and environmental fields[25,27,154,155].

LCA are sensitive to the uncertainties in the underlying data[156]. Previous studies showed that understanding the uncertainty in LCA is at the central importance when interpreting the results. Qin et al., analyzed the uncertainty distributions of the Life Cycle Inventory database[157]. Sillis et al., conducted quantitative uncertainty analysis of LCA for algal biofuel production[158]. Henderson et al., evaluated the sensitivity of the USEtox fate model to the chemical properties such as K_{ow} and k_{degW}, but the uncertainty of the overall fate model in USEtox, as well as the potential to reduce the uncertainty are still unknown.

Due to the massive number of chemicals exist in the current regulatory databases, such as Chemical Abstract Service (CAS) and European Chemicals Agency (ECHA)[15,90], there are demands to conduct LCA in timely manner with reliable accuracy when there

are data gaps[36,159]. Machine learning techniques, opens up new opportunities for LCA practitioners to address data gap when there is only a little information available[160]. Machine learning has the advantage of extracting complex relationship between the predictors and the target values. Researches using machine learning to predict values such as chemical toxicity, bioactivities are pronounced in the area of Quantitative structure–activity relationship (QSAR)[103,161]. In the field of LCA, previous studies have used machine learning methods to estimate the characterized results of organic chemicals in few impact categories, taking molecular structure as inputs[28,162]. Although these models demonstrated the ability of using machine learning to help LCA studies, their model performances sometimes suffer from the problems that the intermediate steps in life cycle impact assessment could not be estimated by molecular structure very well. Moreover, whether using machine learning model is always better than the conventional proxy methods in LCA is still unclear.

In this study, we seek to answer the question that whether replacing the current default proxy methods for chemical properties with machine learning models are always improving the accuracy of impact assessment. We demonstrate it by predicting the chemical properties to estimate one of the intermediate parameters in LCA, the Fate Factor (FF). The data requirements to run the USEtox fate model were assessed. The importance of the chemical properties in terms of their contribution to the USEtox's FFs was evaluated through Global Sensitivity Analysis (GSA). Artificial Neural Network (ANN) and Random Forest (RF) based predictive models were developed to predict these chemical properties, depends on the size of training data. The uncertainties of the default proxy methods and the newly developed machine learning methods were assessed by

comparing the predicted values with experimental data, and the best practice methods were recommended. The uncertainty range of the USEtox FFs were evaluated using Monte Carlo Simulation (MCS).

B. Materials and Methods

The Chemical Fate Model in USEtox. The fate model in USEtox *v2.01* is a multimedia transport and transformation model. It contains many environmental compartments including household indoor air, occupational indoor air, urban air, continental rural air, continental freshwater, continental sea water, continental agricultural soil, continental natural soil and crop residues. The fate model also contains urban, continental and global level as the geographic scale[12]. As a case study, we selected the freshwater compartment and North America continent as the target environmental compartment and geographic scale.

Previous study has evaluated the sensitivity of biodegradation rates in water to the fate model in USEtox [98]. To evaluate the importance of the chemical properties to USEtox, this study conducted Global Sensitivity Analysis (GSA) and compared the contribution of different chemical properties: k_{degW}, K_{ow}, K_{oc}, Sol_{25} and $Pvap_{25}$. In contrast to local sensitivity analysis, where a small perturbation to single model input is studied, GSA seeks to understand the contribution of all model inputs altogether[163]. Cucurachi et al. pointed out the importance of understanding the sensitivity between the results of LCAs and their input parameters, and illustrated how to use GSA to examine the contribution of these paramters[164]. This study adapted three methods for GSA described in previous studies: Kolmogorov-Smirnov Distance Beta (KS) method,

Borgonovo Delta (δ) method and Kuiper Discrepancy Kappa (κ) method[165–167]. These three methods all considered the whole variation range of model inputs. The GSA results can be found in the section 3.1 of this study.

Data Collection and Machine Learning Model Development. The training data to develop the machine learning models for each endpoint in this study are all collected from the PhysProp database, which is embedded in the EPIsuite tool[168]. EPA also provide an online dashboard to retrieve the experimental data of chemical properties[169]. The chemical SMILEs that represent molecular structure were collected from PubChem database[170]. This study focuses on non-ionized organic chemicals. Inorganic chemicals, ionized organic chemicals, as well as the chemicals that can't find SMILEs were removed from the dataset. The final dataset collected in this study for each chemical property can be found in the supporting information.

Molecular structural descriptors were calculated using Python packages rdkit and mordred[123,124]. These two packages together can provide more than 2,200 molecular descriptors, including basic physicochemical properties and autocorrelation descriptors[124]. Large amount of descriptors could lead to overfitting problem, in which the model would perform significantly better on the training dataset, but much worse on the testing dataset.[103,112] To avoid this, and to extract more meaningful subset of molecular descriptors, two steps feature selection algorithm was used in this study. In the first step, a filter-based feature selection method was firstly used to drop descriptors that have variance lower than 5 across all chemicals for this property. Then, the first descriptor in a pair that has correlation higher than 0.95 was dropped. In the second step,

a tree-based model was used to evaluate the contribution of each remaining descriptor to the chemical property, and only the descriptors that contributed above the average contribution were used as the final inputs to train the ANN models[171]. The feature selection process was conducted using Python package sklearn (version 0.2). The final descriptors, and the computer code for feature selection can be found in the supporting information.

In this study, we used fully connected Artificial Neural Networks (ANNs) and Random Forest Regressor (RF) as the modeling basis. For the endpoints that have large amount of training data, previous studies have shown that ANN model can produce better performance[19,172]. For endpoint that doesn't have large enough of training data, like k_{degW}, Random Forest is better since overfitting problem is less likely to occur[173].

ANN is a model structure can be used to approximate the relationship between inputs and outputs at higher dimensions. Firstly used in early 1980s, ANN nowadays have been applied in many products in the field of Artificial Intelligence[174–176], as well as in the field of chemoinformatic and (Quantitative Structure-Activity Relationship) QSAR[46,161,177]. Random forest (regressor) model (RF) is essentially a group of decision tree, and with bagging and bootstrapping techniques when taking inputs for different trees[178].

In our study, the ANNs and RF were developed with Python packages of Tensorflow and Keras[128,129]. For both ANNs and RF models, 10% of the entire dataset for each endpoint was randomly selected as the testing dataset, and the average performance on the validation datasets in five-fold-cross-validation was used to evaluate the performance of the hyper-parameters. For ANNs, the hyper-parameters are: the number

of hidden layer(s), the number of hidden neuron(s) in each layer, the regularization factor and the type of activation function. For the RF model, the optimized parameter was the number of decision tree in the forest. The best hyper-parameters were used to create the final models using the entire dataset except the testing dataset, for both ANN and RF. The final model performances were reported on the testing dataset.

Model Errors and the Best Practice Methods. One of the major goals of this study is to compare the errors made by the machine learning models as well as the default proxy methods. To do so, for each model, the absolute errors between the value predicted by both proxy methods (machine learning, blue and default, orange) were compared with the experimental datasets, as shown in Equation 1:

$$E_i^{n,k} = P_i^{n,k} - Exp_i^n \quad (1)$$

where $E_i^{n,k}$ is the prediction error using method *k* (*proxy* or *ml*) for chemical property *n* and for chemical *i*; $P_i^{n,k}$ is the predicted value using method *k* for chemical *i* on property *n*; Exp_i^n is the experimental value for chemical *i* on property *n*, serving as the ground truth for the predicted value to compare with.

For each chemical property, the testing chemicals, which are 10% of the entire collected data on each property, were used to characterize the errors of the default proxy methods (E_i^{proxy}) and the machine learning models (E_i^{ml}). Therefore, the testing data for the same chemical property were the same between different methods, but might be different between different chemical property. The distributions of the errors were fitted to normal distributions, so that the mean (μ^k) and the standard deviation (σ^k) of method *k* can be estimated and compared. Between the machine learning models and the default

proxy methods, the ones that generated the smaller errors (μ^k closer to zero and/or smaller σ^k) were selected as the "best practice methods" and used to generate new Characterization Factors.

Characterizing Uncertainty in the Fate Factor. To evaluate the uncertainty in the FF caused by the errors introduced by the proxy methods suggested by USEtox, Monte Carlo Simulation (MCS) was used in this study. Since the uncertainty ranges we characterized (in section 2.3) were associated with each chemical property, they are irrelevant to chemicals. Therefore, we randomly selected a chemical Tribufos (78-48-8) as an example to run MCS. We run USEtox model 10,000 times. In each time, the values for each chemical property we built model for were predicted, and the prediction errors were sampled from the distribution curves we characterized in section 2.3. Therefore, the values used to run USEtox during MCS were calculated as in Equation 2:

$$I_i^{n,k} = P_i^{n,k} - E_i^{n,k} \quad (2)$$

where I_i^k is the input values we used to run USEtox in MCS for chemical *i*, property *n* using method *k* (proxy or ml); $P_i^{n,k}$ is the predicted value generated by method *k* for the same chemical *i* on property *n*; $E_i^{n,k}$ is the prediction error we sampled from the distribution curves, generated by the mean (μ^k) and the standard deviation (σ^k) for method *k*, property *n* and chemical *i*, which we estimated in section 2.3.

To demonstrate how the uncertainty of FFs of USEtox can be reduced by using machine learning techniques, the MCS was conducted twice for the same chemical using the chemical properties generated by the best practice methods as well as the default

proxy methods. The uncertainties in the FFs using these two predictive methods are presented in the Results and Discussion section.

C. Results and Discussion

Sensitivity of the USEtox model to chemical properties. The sensitivity of the USEtox v2.01 fate model to the variation in the input chemical properties were estimated using GSA, and the results are presented in Table 1. The results were estimated for 1 kg of 4-nitroaniline in fresh water compartment. The experimental data were used when available, and the default proxy methods were used to fill in the missing input data. As the results indicated in Table 1, in all sensitivity analysis methods, k_{degW} (the biodegradation rate in water) shows the highest importance (in KS:0.73, in Delta: 0.97 and in K: 0.91). The contribution of the other chemical properties such as K_{ow}, Sol_{25} and K_{oc} were about 13 to 17 times smaller than k_{degW}, respectively. Since this study is focusing on the emission to fresh water compartment, the remaining chemical properties: k_{degA}, k_{degSd}, k_{degSl} (biodegradation rate in air, sediment and soil) have no contribution to the fate factors in this case.

Table 1. The sensitivity of the USEtox FFs for emission to fresh water compartment to chemical properties using three different global sensitivity analysis method. KS (Kolmogorov-Smirnov Distance Beta), Delta (Borgonovo Delta) and K (Kuiper Discrepancy Kappa). The numbers indicate a score for the importance of chemical properties by different methods. The scores are not necessary summing up to one.

	KS	δ	κ
k_{degW}	0.73	0.97	0.91
K_{OW}	0.05	0.06	0.07
Sol_{25}	0.05	0.06	0.07
$Pvap_{25}$	0.05	0.07	0.07
K_{oc}	0.04	0.07	0.07
k_{degA}	0.00	0.00	0.00
k_{degSd}	0.00	0.00	0.00

k_{degSl}	0.00	0.00	0.00

Machine learning models' performances. The performance of the machine learning models, for k_{degW} and K_{oc}, developed based on the training data collected in this study are presented in Figure 1. The performances of the other models are presented in the supporting information. The statistics of these models (R^2 and the number of training and testing data) are presented in Table 2.

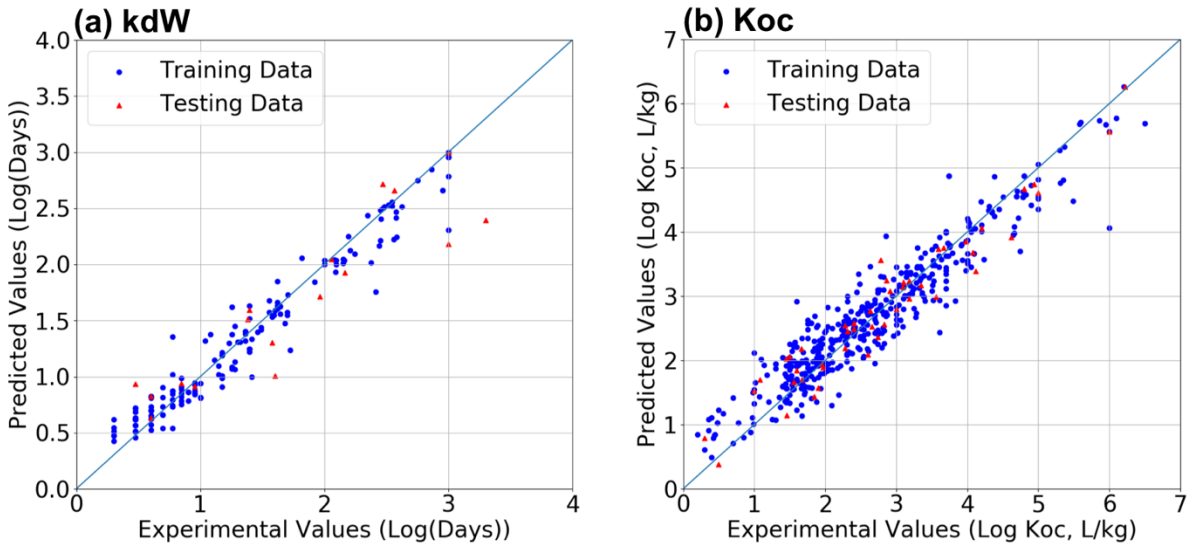

Figure 1. The performances of the machine learning model for k_{degW} (a, developed by Random Forest) and K_{oc} (b, developed by ANN) on the training and testing data.

The machine learning models showed good performances on the testing chemicals for most chemical properties. The model with the highest R^2 values on testing chemicals is the K_{oc} and $Pvap_{25}$ models, with R^2 0.83 for both, followed by the k_{degW} model, with R^2 0.81 on the testing data. The model with the lowest R^2 is the one for K_{ow} (0.67 on the

testing data). Due to the limited number of the experimental data can be found for k_{degW}, the machine learning model for it was developed using Random Forest model. The models for the other endpoints are based on ANN.

Table 2. The R^2 values of the machine learning models for chemical properties on the training and testing data, along with the number of chemicals in the training and testing data for each model.

	R^2 on Training Data	R^2 on Testing Data	Number of Training Data	Number of Testing Data
k_{degW}	0.97	0.81	158	17
K_{oc}	0.87	0.83	441	48
K_{ow}	0.89	0.67	2265	251
Sol_{25}	0.87	0.76	2172	241
Vap_{25}	0.91	0.83	1425	158

Comparing the default proxy method and machine learning models. To decide the best practice methods to estimate the inputs to the fate model in USEtox, the errors $E_i^{n,k}$ in Equation 1 of these two methods (machine learning and default) on groups of chemicals that the experimental data are known were estimated in this study.

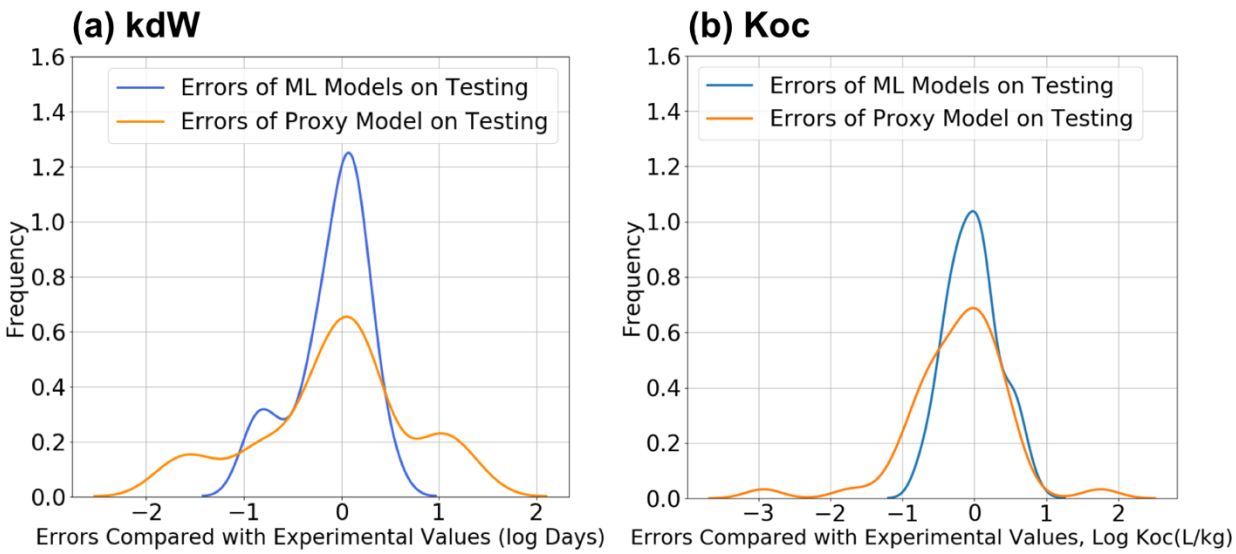

Figure 2. The absolute errors between the default proxy methods and the experimental value (orange), and between the machine learning proxy methods and the experimental values (blue), for k_{degW} and $Pvap_{25}$.

Figure 2 shows the comparison for chemical properties k_{degW} and $Pvap_{25}$. The statistics of the distributions are shown in Table 3. The experimental value used as the baseline were the chemicals in the testing dataset for each endpoint. As a result, the machine learning models for k_{degW}, K_{ow}, K_{oc} and Sol_{25} showed improvement to the default proxy methods. For these three endpoints compared with default proxy methods, the standard deviation (σ) of prediction errors was all reduced, from 6.02 to 2.34 days for k_{degW}, and from 4.17 to 2.57 L/kg for K_{oc}, from 26.3 to 16.21 L/L for K_{ow}, and from 64.56 to 9.77 mg/L for Sol_{25}. The machine learning models for $Pvap_{25}$ didn't show satisfied improvement, the mean error and the error standard deviation of the machine learning model was 1.35 and 22.59 Pa, respectively, while the default proxy method recommend by USEtox achieved 1.12 and 25.11 Pa, respectively. Therefore, in this study, we selected the machine learning models for k_{degW}, K_{oc}, K_{ow} and Sol_{25} and the default proxy method

for $Pvap_{25}$ as the best practice method to estimate the inputs to the USEtox fate model when experimental data is missing.

Table 3. The mean value (μ) and standard deviation (σ) for the default proxy methods and machine learning models for each endpoint compared with experimental values in the testing dataset

	Mean Errors (μ)		standard deviation (σ)		Machine Learning Model as the Best Practice Method?
	default proxy	machine learning	default proxy	machine learning	
k_{degW}	-0.07	-0.1	0.78	0.37	✓
K_{oc}	-0.09	0	0.62	0.41	✓
K_{ow}	0.43	0.31	1.42	1.21	✓
Sol_{25}	-0.07	0.03	1.81	0.99	✓
$Pvap_{25}$	0.05	0.13	1.4	1.35	×

Comparing the default proxy method and machine learning based method. To decide the best practice methods to estimate the inputs to the fate model in *USEtox*. The errors of these two proxy methods on groups of chemicals that the experimental data are known are presented in this study.

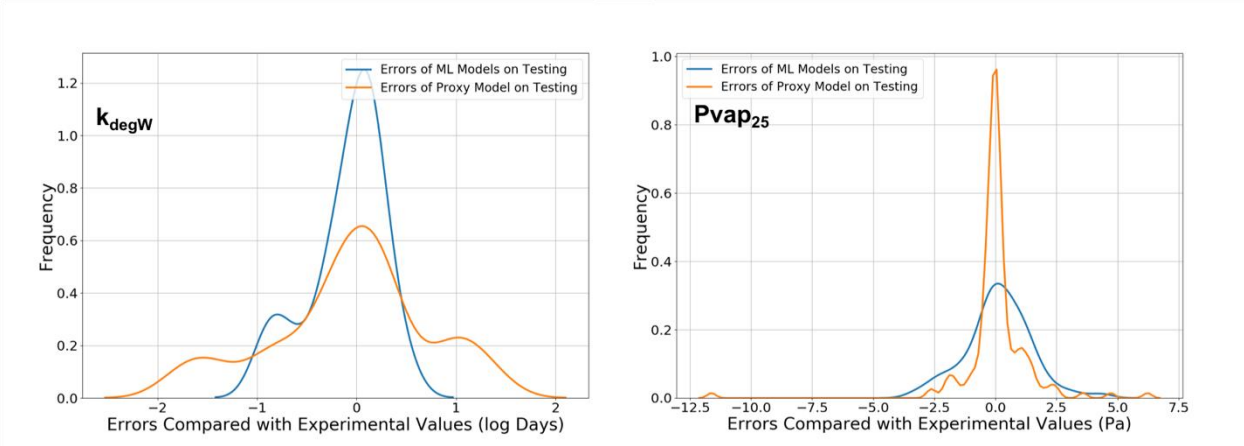

Figure 3. The absolute errors between the default proxy methods and the experimental value (orange), and between the machine learning proxy methods and the experimental values (blue), for k_{degW} and $Pvap_{25}$.

Figure 3 shows the comparison for the chemical properties k_{degW} and $Pvap_{25}$. For each endpoint, the absolute errors between the value predicted by both proxy methods (machine learning, blue and default, orange) were compared with the experimental datasets. To characterize the uncertainties of these two methods, the errors are approximated using normal distribution. The mean values and standard deviations for each endpoint were estimated. The statistics of the distributions are shown in Table 4. The experimental value used as the baseline are the chemicals in the testing dataset for each endpoint. As the results of comparison with experimental data for each endpoint, the machine learning models for k_{degW}, K_{ow}, K_{oc} and Sol_{25} showed improvement to the default proxy methods. For these three endpoints compared with default proxy methods, the standard deviation of prediction errors were all reduced, from 0.78 0.37 log days for k_{degW}, and from 0.62 to 0.41 log L/kg for K_{oc}, from 1.42 to 1.21 log L/L for K_{ow}, and from 1.81 to 0.99 log mg/L for Sol_{25}. The machine learning models for $Pvap_{25}$ didn't show satisfied improvement, the mean error and the error standard deviation of the machine learning model was 0.13 and 1.35 log Pa, respectively, while the default proxy

method recommend by *USEtox* achieved 0.05 and 1.40 log Pa, respectively. Therefore, in this study, I selected the machine learning models for k_{degW}, K_{oc}, K_{ow} and Sol_{25} and the default proxy method for $Pvap_{25}$ as the best practice method to estimate the inputs to the *USEtox* fate model when experimental data is missing.

Table 4 The mean value (μ) and standard deviation (σ) for the default proxy methods and machine learning models for each endpoints compared with experimental values in the testing dataset

	$\mu_{default\ proxy}$	$\sigma_{default\ proxy}$	$\mu_{machine\ learning}$	$\sigma_{machine\ learning}$
k_{degW}	-0.07	0.78	-0.10	0.37
K_{oc}	-0.09	0.62	0.00	0.41
K_{ow}	0.43	1.42	0.31	1.21
Sol_{25}	-0.07	1.81	0.03	0.99
$Pvap_{25}$	0.05	1.40	0.13	1.35

Uncertainty of the Fate Factors. MCS was used to estimate the uncertainty of the FFs, using both the best practice methods and the default proxy methods to estimate the inputs for *tribufos* (CAS: 78-48-8). For each chemical property (k_{degW}, K_{oc}, K_{ow}, Sol_{25} and $Pvap_{25}$), the means and the standard deviations in Table 4 were used to sample the inputs, from normal distributions, to the USEtox fate model for 10,000 times. Figure 4 shows the density of the results of 10,000 times of MCS. The blues bins were generated using the best practice methods (defined in Table 4), and the reds bins were generated using the default proxy methods provided by USEtox. As the Figure 4 indicates, the mean values of the FFs using these two methods were close for *tribufos* (CAS: 78-48-8). The uncertainty ranges of the FFs estimated using the default proxy methods were from about 1.58 to 630.96 kg/kg·days^{-1}, while the uncertainty reduced to the range of 9.54 to 380.19 kg/kg·days^{-1} when the best practice methods were used.

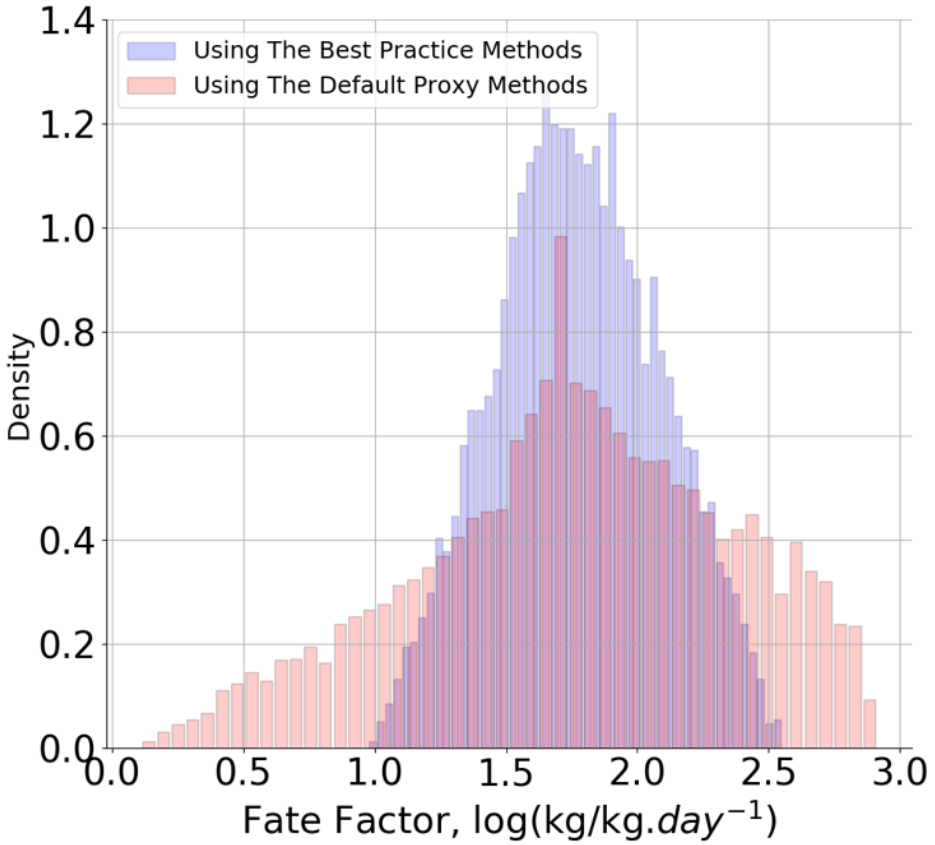

Figure 4. The distributions of 10,000 times MCS for the USEtox FF, using the best practice methods (blue), as well as the default proxy methods (red) to estimate k_{degW}, K_{oc}, K_{ow}, Sol_{25} and $Pvap_{25}$ for tribufos (CAS: 78-48-8).

USEtox is one of the most used life cycle impact assessment models to estimate the human health and ecosystem impact of chemicals. The methodologies used in the USEtox fate model are the results of the scientific consensus of United Nations Environment Programme/Society of Environmental Toxicology and Chemistry (UNEP/SETAC). The CFs can be regarded as accurate when the inputs are in high quality. However, the uncertainty of the USEtox CFs introduced by the uncertainties in the chemical properties as model inputs have never been studied in current literature. The

object of this chapter is to improve the accuracy of the outputs of the USEtox model by reducing the uncertainties in the inputs, instead of validating the correctness of the model.

The results of this study show that using machine learning based model can significantly improve the uncertainty in the FFs, as indicated in Figure 4, compared with using the default proxy methods, but not always. The default proxy methods recommended by USEtox were mostly based on the relationship between physicochemical properties. For example, estimating K_{oc} using K_{ow}, or using chemical half-live time in low resolution to estimate the biodegradation rate. These methods, although have been well-established and peer-reviewed in pervious literature, do introduced considerable uncertainties when the input chemicals become more complex.

Machine learning based models have the advantage of fully utilizing the existing experimental data. As the computational techniques advance and more experimental data become available in variety formats, machine learning models can be developed with more training data nowadays, which results to improvements in the model performances. It is necessary to point out that the quality machine learning base models, due to this nature mentioned above, relies on the quality of their training data (so called "garbage in, garbage out")[179].

This study aims to resolve the challenge to conduct LCA at a screening level, when only a little information about the chemical is known. When the experimental data is missing and no EFs and FFs can be found, chemical structural information can be used as an effective predictor to estimate the model parameters to calculate FF. The machine learning models in this study demonstrated that the intermediate parameters in impact assessment, like the FF, can be modeled by using the reliable inputs generated by

machine learning models, and the only required information is the molecular structure. Given the millions of existing organic chemicals registered in regulatory databases[15,90], the outcomes of this study help reduce the cost and time to run LCA for organic chemicals.

D. Acknowledgement

The author thanks Prof. Sangwon Suh for his helps over the development of this chapter. The author also want to thanks Prof. Dingsheng Li for his help in understanding *USEtox* model, and Dr. Yuwei Qin for the help in sensitivity analysis. This publication was developed under Assistance Agreement No. 83557901 awarded by the U.S. Environmental Protection Agency to the University of California, Santa Barbara. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication.

Appendix I: Supporting Information for Chapter II

The List of Training Chemicals. The training chemicals were collected from Ecoinvent 3.01 life-cycle inventory database. It contains 166 organic chemicals. The number of chemicals in each dataset and their use are reported in Table S1. The list of collected chemicals are presented in Table S2. The full descriptors and chemical name can be found in the Excel file of Supporting Information.

Table S1. Number of chemicals in the training, validation and testing data.

	Number of Chemicals	Notes
Training data	146	Used to train the ANN model
Validation data	16	Used to report model performance during training and to tune hyper-parameters
Test data	10	Used to report the final performance of models

Table S2. The list of organic chemicals we used in this study. Along with the SMILEs used to calculate molecular descriptors.

Name	SMILEs	Name	SMILEs
1-propanol	CCCO	glyoxal	C(=O)C=O
1,1-difluoroethane	C(C(OC(F)F)(F)F)(F)Cl	hexafluoroethane	C(C(F)(F)F)(F)(F)F
2-butanol	CCC(C)O	hydroquinone	c1cc(ccc1O)O
2-methyl-2-butanol	CCC(C)(C)O	imidazole	c1cnc[nH]1
2-nitroaniline	c1ccc(c(c1)N)[N](=O)[O]	isobutyl acetate	CC(C)COC(=O)C
2, 4-dichlorophenol	Clc1cc(Cl)c(O)cc1	isohexane	CCCC(C)C
2, 4-dichlorotoluene	c1cc(ccc1CCl)Cl	isopropanol	CC(C)O
3-methyl-1-butyl acetate	CC(C)CCOC(=O)C	isopropyl acetate	CC(C)OC(=O)C
4-methyl-2-pentanone	CC(C)CC(=O)C	isopropylamine	CC(C)N
4-tert-butylbenzaldehyde	CC(C)(C)c1ccc(cc1)C=O	lactic acid	CC(C(=O)O)O
4-tert-butyltoluene	Cc1ccc(cc1)C(C)(C)C	maleic anhydride	C1=CC(=O)OC1=O

acetaldehyde	CC=O	melamine	c1(nc(nc(n1)N)N)N
acetanilide	C/C(=N\c1ccccc1)/O	meta-phenylene diamine	NN
acetic acid	CC(=O)O	methacrylic acid	CC(=C)C(=O)O
acetic anhydride	CC(=O)OC(=O)C	methane sulfonic acid	CS(=O)(=O)O
acetoacetic acid	CC(=O)CC(=O)O	methanol	CO
acetone	CC(=O)C	methyl acrylate	COC(=O)C=C
acetyl chloride	CC(=O)Cl	methyl ethyl ketone	CCC(=O)C
acetylene	C#C	methyl formate	COC=O
acrolein	C=CC=O	methyl iodide	CI
acrylic acid	C=CC(=O)O	methyl tert-butyl ether	CC(C)(C)OC
adipic acid	C(CCC(=O)O)CC(=O)O	methyl-3-methoxypropionate	COCCC(=O)OC
allyl chloride	C=CCCl	methylamine	CN
alpha-naphthol	c1ccc2c(c1)cccc2O	methylchloride	CCl
alpha-picoline	Cc1ccccn1	methylcyclohexane	CC1CCCCC1
aniline	c1ccc(cc1)N	N-methyl-2-pyrrolidone	CN1CCCC1=O
anthranilic acid	c1ccc(c(c1)C(=O)O)N	N, N-dimethylformamide	CN(C)C=O
benzal chloride	c1ccc(cc1)C(Cl)Cl	naphthalene sulfonic acid	c1ccc2c(c1)cccc2S(=O)(=O)O
benzaldehyde	c1ccc(cc1)C=O	nitrobenzene	c1ccc(cc1)[N](=O)[O]
benzyl alcohol	c1ccc(cc1)CO	o-aminophenol	c1ccc(c(c1)N)O
benzyl chloride	c1ccc(cc1)CCl	o-chlorobenzaldehyde	c1ccc(c(c1)C=O)Cl
bisphenol A	CC(C)(c1ccc(cc1)O)c2ccc(cc2)O	o-chlorotoluene	Cc1ccccc1Cl
boron trifluoride	B(F)(F)F	o-cresol	Cc1ccccc1O
bromopropane	CCCBr	o-nitrophenol	c1ccc(c(c1)[N](=O)[O])O
butane	CCCC	ortho-phenylene diamine	NN
butane-1, 4-diol	CS(=O)(=O)OCCCCOS(=O)(=O)C	p-chlorophenol	c1cc(ccc1O)Cl
butyl acetate	CCCCOC(=O)C	p-nitrophenol	c1cc(ccc1[N](=O)[O])O
butyl acrylate	CCCCOC(=O)C=C	p-nitrotoluene	Cc1ccc(cc1)[N](=O)[O]
carbon tetrachloride	C(Cl)(Cl)(Cl)Cl	para-phenylene diamine	c1cc(ccc1N)N
chloroacetic acid	C(C(=O)O)Cl	pentaerythritol	C(C(CO)(CO)CO)O

chloroacetyl chloride	C(C(=O)Cl)Cl	pentane	CCCCC
chlorodifluoromethane	C(F)(F)Cl	perfluoropentane	C(C(C(F)(F)F)(F)F)(C(C(F)(F)F)(F)F)(F)F
chloromethyl methyl ether	COCCl	phenol	c1ccc(cc1)O
chloronitrobenzene	c1ccc(c(c1)[N](=O)[O])Cl	phenyl acetic acid	c1ccc(cc1)CC(=O)O
chloropropionic acid	CC(C(=O)O)Cl	phenyl isocyanate	c1ccc(cc1)N=C=O
cumene	CC(C)c1ccccc1	phosgene	C(=O)(Cl)Cl
cyanoacetic acid	C(C#N)C(=O)O	phosphorous chloride	P(Cl)(Cl)Cl
cyanogen chloride	C(#N)Cl	phosphorus pentachloride	P(Cl)(Cl)(Cl)(Cl)Cl
cyanuric chloride	c1(nc(nc(n1)Cl)Cl)Cl	phosphoryl chloride	O=P(Cl)(Cl)Cl
cyclohexane	C1CCCCC1	phthalic anhydride	c1ccc2c(c1)C(=O)OC2=O
cyclohexanol	C1CCC(CC1)O	phthalimide	c1ccc2c(c1)C(=NC2=O)O
cyclohexanone	C1CCC(=O)CC1	piperidine	C1CCNCC1
dichloromethane	C(Cl)Cl	polyacrylamide	c1ccc(cc1)/C=C(/C(=O)N)\N
diethanolamine	C(CO)NCCO	propanal	CCC=O
diethyl ether	CCOCC	propionic acid	CCC(=O)O
diethylene glycol	C(COCCO)O	propyl amine	CCCN
dimethyl ether	COC	propylene	CC=C
dimethyl malonate	COC(=O)CC(=O)OC	propylene glycol	C[C@H](CO)O
dimethyl sulfate	COS(=O)(=O)OC	propylene oxide	CC1CO1
dimethyl sulfide	CSC	pyrazole	c1c[nH]nc1
dimethyl sulfoxide	CS(=O)C	sodium methoxide	C[O][Na]
dimethylacetamide	CC(=O)N(C)C	styrene	C=Cc1ccccc1
dimethylamine	CNC	tert-butyl amine	CC(C)(C)N
dioxane	C1COCCO1	tetrachloroethylene	C(=C(Cl)Cl)(Cl)Cl
dipropyl amine	CCCNCCC	tetraethyl orthosilicate	CCO[Si](OCC)(OCC)OCC
dipropylene glycol monomethyl ether	CC(CO)OCC(C)OC	tetrafluoroethane	C(C(F)(F)F)F
DTPA	C(CN(CC(=O)O)CC(=O)O)N(CCN(CC(=O)O)CC(=O)O)CC(=O)O	tetrahydrofuran	C1CCOC1
EDTA	C(CN(CC(=O)O)CC(=O)O)N(CC(=O)O)CC(=O)O	toluene	Cc1ccccc1
epichlorohydrin	C1C(O1)CCl	trichloroacetic acid	C(=O)(C(Cl)(Cl)Cl)O
ethyl acetate	CCOC(=O)C	trichloroborane	B(Cl)(Cl)Cl

ethyl benzene	CCc1ccccc1	trichloroethylene	C(=C(Cl)Cl)Cl
ethylamine	CCN	trichloromethane	C(Cl)(Cl)Cl
ethylene bromide	C(CBr)Br	trichloropropane	C(C(CCl)Cl)Cl
ethylene carbonate	C1COC(=O)O1	triethyl amine	CCN(CC)CC
ethylene dichloride	C(CCl)Cl	trifluoroacetic acid	C(=O)(C(F)(F)F)O
ethylene glycol diethyl ether	CCOCCOCC	trifluoromethane	C(F)(F)F
ethylene glycol dimethyl ether	COCCOC	trimesoyl chloride	c1c(cc(cc1C(=O)Cl)C(=O)Cl)C(=O)Cl
ethylene glycol monoethyl ether	CCOCCO	trimethyl borate	B(OC)(OC)OC
ethylene oxide	C1CO1	trimethylamine	CN(C)C
ethylenediamine	C(CN)N	vinyl acetate	CC(=O)OC=C
formic acid	C(=O)O	vinyl chloride	C=CCl
glycerine	C(C(CO)O)O	vinyl fluoride	C=CF
glycine	C(C(=O)O)N	xylene	Cc1ccccc1C

The List of Molecular Descriptors Used in This Study. The molecular descriptors we used in this study were generated through the software *Dragon 7*. We used this software and calculated 3,839 molecular descriptors, including constitutional, ring, adjacency and other types of descriptors. We applied the filter-based feature-selection method and reduced the number of descriptors to 58. The full list of the reduced descriptors and their full name is showing in Table S1. We also used Principle Component Analysis (PCA) and extracted 60 features, which preserved 95% variance in all descriptors calculated by *Dragon 7*. Figure S1 shows the number of extracted descriptors by PCA against the cumulative variance preserved in the full descriptor set.

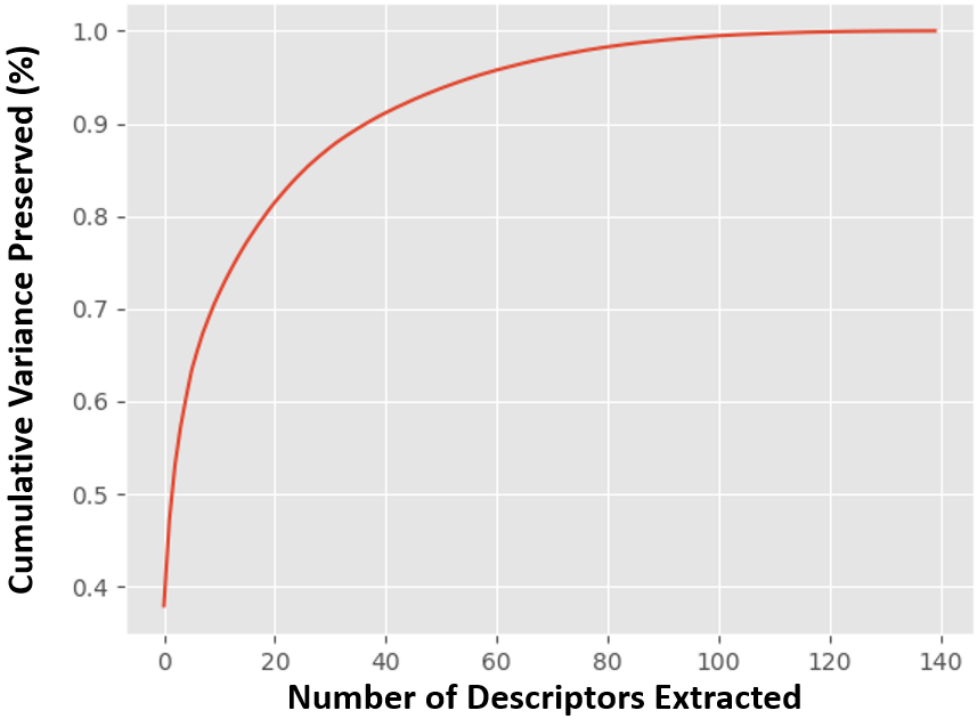

Figure S1. The number of descriptors extracted by PCA against the cumulative variance preserved by the corresponding descriptors.

Table S3. List of molecular descriptors produced by the filter-based feature selection method

Descriptors Abbreviation	Descriptors Full Name	Descriptor Category
MW	molecular weight	Constitutional indices
AMW	average molecular weight	Constitutional indices
nBM	number of multiple bonds	Constitutional indices
RBN	number of rotatable bonds	Constitutional indices
nF	number of Fluorine atoms	Constitutional indices
N%	percentage of N atoms	Constitutional indices
O%	percentage of O atoms	Constitutional indices
D/Dtr05	distance/detour ring index of order 5	Ring descriptors
D/Dtr10	distance/detour ring index of order 10	Ring descriptors
MAXDP	maximal electrotopological positive variation	Topological indices
Psi_i_A	intrinsic state pseudoconnectivity index - type S average	Topological indices
Yindex	Balaban Y index	Information indices
CIC4	Complementary Information Content index (neighborhood symmetry of 4-order)	Information indices
CIC5	Complementary Information Content index (neighborhood symmetry of 5-order)	Information indices
VR1_D/Dt	Randic-like eigenvector-based index from distance/detour matrix	2D matrix-based descriptors
SpDiam_B(m)	spectral diameter from Burden matrix weighted by mass	2D matrix-based descriptors

ATSC2m	Centred Broto-Moreau autocorrelation of lag 2 weighted by mass	2D autocorrelations
ATSC1p	Centred Broto-Moreau autocorrelation of lag 1 weighted by polarizability	2D autocorrelations
GATS6m	Geary autocorrelation of lag 6 weighted by mass	2D autocorrelations
GATS7s	Geary autocorrelation of lag 7 weighted by I-state	2D autocorrelations
P_VSA_LogP_1	P_VSA-like on LogP, bin 1	P_VSA-like descriptors
P_VSA_LogP_2	P_VSA-like on LogP, bin 2	P_VSA-like descriptors
P_VSA_LogP_8	P_VSA-like on LogP, bin 8	P_VSA-like descriptors
P_VSA_MR_3	P_VSA-like on Molar Refractivity, bin 3	P_VSA-like descriptors
P_VSA_MR_5	P_VSA-like on Molar Refractivity, bin 5	P_VSA-like descriptors
P_VSA_MR_7	P_VSA-like on Molar Refractivity, bin 7	P_VSA-like descriptors
P_VSA_s_1	P_VSA-like on I-state, bin 1	P_VSA-like descriptors
P_VSA_s_3	P_VSA-like on I-state, bin 3	P_VSA-like descriptors
P_VSA_ppp_D	P_VSA-like on potential pharmacophore points, D	P_VSA-like descriptors
SpDiam_EA(dm)	spectral diameter from edge adjacency mat. weighted by dipole moment	Edge adjacency indices
SM14_AEA(dm)	spectral moment of order 14 from augmented edge adjacency mat. weighted by dipole moment	Edge adjacency indices
SM15_AEA(dm)	spectral moment of order 15 from augmented edge adjacency mat. weighted by dipole moment	Edge adjacency indices
SM02_AEA(ri)	spectral moment of order 2 from augmented edge adjacency mat. weighted by resonance integral	Edge adjacency indices
SM04_AEA(ri)	spectral moment of order 4 from augmented edge adjacency mat. weighted by resonance integral	Edge adjacency indices

SM06_AEA(ri)	spectral moment of order 6 from augmented edge adjacency mat. weighted by resonance integral	Edge adjacency indices
SM10_AEA(ri)	spectral moment of order 10 from augmented edge adjacency mat. weighted by resonance integral	Edge adjacency indices
nCp	number of terminal primary C(sp3)	Functional group counts
nCs	number of total secondary C(sp3)	Functional group counts
H-046	H attached to C0(sp3) no X attached to next C	Atom-centred fragments
H-047	H attached to C1(sp3)/C0(sp2)	Atom-centred fragments
H-051	H attached to alpha-C	Atom-centred fragments
H-052	H attached to C0(sp3) with 1X attached to next C	Atom-centred fragments
SssO	Sum of ssO E-states	Atom-type E-state indices
CATS2D_02_DL	CATS2D Donor-Lipophilic at lag 02	CATS 2D
CATS2D_02_AA	CATS2D Acceptor-Acceptor at lag 02	CATS 2D
CATS2D_02_AL	CATS2D Acceptor-Lipophilic at lag 02	CATS 2D
CATS2D_03_AL	CATS2D Acceptor-Lipophilic at lag 03	CATS 2D
CATS2D_05_AL	CATS2D Acceptor-Lipophilic at lag 05	CATS 2D
CATS2D_04_LL	CATS2D Lipophilic-Lipophilic at lag 04	CATS 2D
T(N..Cl)	sum of topological distances between N..Cl	2D Atom Pairs
T(O..F)	sum of topological distances between O..F	2D Atom Pairs

T(O..Cl)	sum of topological distances between O..Cl	2D Atom Pairs
T(F..Cl)	sum of topological distances between F..Cl	2D Atom Pairs
F03[C-O]	Frequency of C - O at topological distance 3	2D Atom Pairs
F03[C-Cl]	Frequency of C - Cl at topological distance 3	2D Atom Pairs
MLOGP2	squared Moriguchi octanol-water partition coeff. (logP^2)	Molecular properties

The Impact Categories in this Study. At this point we are able to estimate six impact categoires for organic chemicals: cumulative energy demand (CED), acidification, global warming, ecoindicator 99, human health, and ecosystem quality. The first three are midpoint impact categories and the latter three are endpoint impact categories. Detailed explanations for each of the endpoints are as follows:

Cumulative energy demand (MJ eq./kg): It is measuring the cradle-to-gate energy consumption to manufacture one kilogram of chemicals. Accumulated through non-renewable (fossil fuel), non-renewable (nuclear), renewable (biomass), renewable (wind, solar, geothermal) and renewable (water) energy

TRACI acidification (molecules of H+ eq./kg): It is measuring the impact on acidification throughout cradle-to-gate product life cycle. This is only the measurement of impact by increasing hydrogen ion without considering the site-specific factors such as the ability of buffering.[180]

Global warming, 100a, IPCC 2007: The impact category of global warming is measuring the global warming potential (GWP) of a chemical, which is the relative effect of a chemical to carbon dioxide on Global Warming. Intergovernmental Panel on Climate Change (IPCC) updated the GWP value for hundreds of chemicals on 2007. The calculation of global warming for a chemical is based on their radiative efficiency and the atmospheric lifetime.[181]

Ecoindicator 99, (I,I): total,total (point/kg): There are many impact categories in LCA and it is difficult to have a meaningful sense from these numbers. This endpoint is designed to assign an overall environmental impact score to products that weighted by the damage to human health, damage to ecosystem quality and the damage to resources

throughout product life cycle. The unit is 'point' as the main purpose is to compare the impact between products and components. It is the damage to individualist and normalization with the individualist weighting.[182]

Impact 2002+, human health, total (DALY/kg): The human health endpoint impact category is the sum of the midpoint categories "human toxicity", "respiratory effect", "ionizing radiation", "ozone layer depletion" and "photochemical oxidation". It is an overall score about how the chemical affect human health from different perspectives.[68]

Impact 2002+, ecosystem quality, total (PDF·m^2·year/kg): The endpoint impact category "ecosystem quality" is the sum of the midpoint categories "aquatic ecotoxicity", "terrestrial ecotoxity", "terrestrial acid/nutr", "land occupation, "aquatic acidification", "aquatic eutrophication" and "water turbined". It is an overall score about how the chemical affect the ecosystem from different perspectives.[68]

The histogram of these six impact categories for the organic chemicals in Ecoinvent v3.01 database are presented in Figure S2.

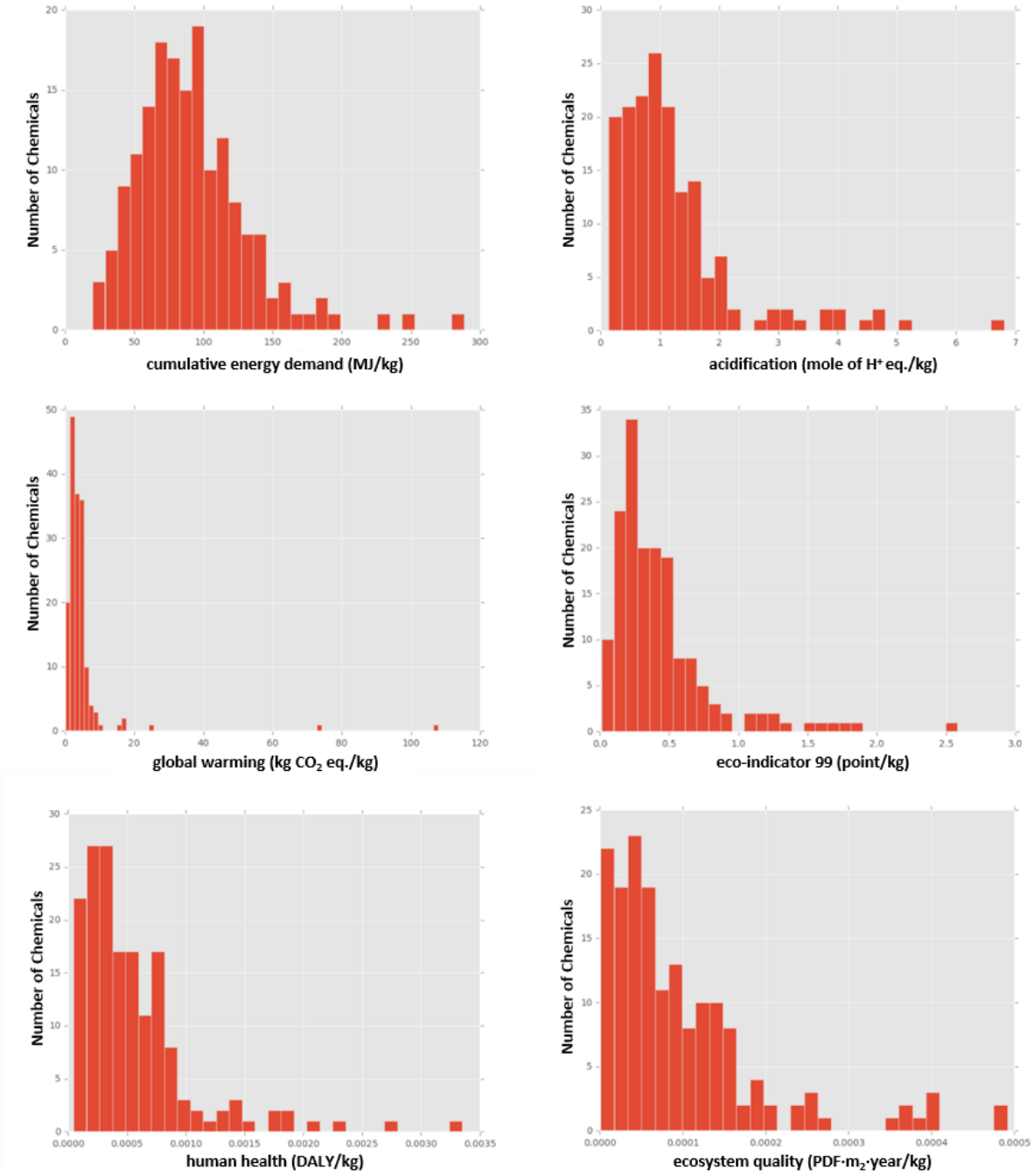

Figure S2. Histogram of the characterized results for the six selected impact categories.

Model Optimization and Development. The performances of all six models that were developed using (1) full descriptors calculated from Dragon 7 (3,839 descriptors), (2) descriptors selected with filter-based method, (3) feature extracted using PCA, and considering 1, 2, 3 hidden layer(s) and 16, 64, 128, 512 hidden neurons in each hidden layers were presented in Table S4 – Table S9. There are 72 different models (6 impact categories, 3 levels of hidden layers and 4 levels of hidden neurons).

Table S4. Validation of the performances for the CED model, developed with different model settings and full descriptors, feature selected descriptors and PCA descriptors as input.

Number of Hidden Layer(s)	Number of Hidden Neuron (s)	Full Descriptors	Feature Selection	PCA
1	16	0.34	0.36	0.38
	64	0.48	0.38	0.44
	128	0.45	0.42	0.51
	512	0.3	0.24	0.38
2	16	0.28	0.46	0.28
	64	0.5	0.45	0.39
	128	0.44	0.36	0.38
	512	0.32	0.33	0.35
3	16	0.14	0.15	0.22
	64	0.12	0.11	0.18
	128	0.02	-0.06	0.09
	512	-0.12	0.03	-0.03

Table S5. Validation of the performances for the acidification model, developed with different model settings and full descriptors, feature selected descriptors and PCA descriptors as input.

Number of Hidden Layer(s)	Number of Hidden Neuron (s)	Full Descriptors	Feature Selection	PCA
1	16	0.33	0.52	0.45
	64	0.42	0.62	0.52
	128	0.4	0.49	0.56
	512	0.51	0.65	0.7
2	16	0.55	0.52	0.58
	64	0.44	0.68	0.69
	128	0.33	0.65	0.75
	512	0.42	0.65	0.74
3	16	0.51	0.68	0.73
	64	0.45	0.59	0.61
	128	0.36	0.42	0.41
	512	0.29	0.41	0.33

Table S6. Validation of the performances for the EI99 model, developed with different model settings and full descriptors, feature selected descriptors and PCA descriptors as input.

Number of Hidden Layer(s)	Number of Hidden Neuron (s)	Full Descriptors	Feature Selection	PCA
1	16	0.21	0.25	0.36
	64	0.4	0.31	0.41
	128	0.42	0.45	0.44
	512	0.45	0.42	0.6
2	16	0.42	0.39	0.63
	64	0.56	0.49	0.66
	128	0.35	0.42	0.6
	512	0.3	0.35	0.44
3	16	0.4	0.39	0.46
	64	0.38	0.41	0.5
	128	0.21	0.36	0.5
	512	0.05	0.29	0.31

Table S7. Validation of the performances for the global warming model, developed with different model settings and full descriptors, feature selected descriptors and PCA descriptors as input.

Number of Hidden Layer(s)	Number of Hidden Neuron (s)	Full Descriptors	Feature Selection	PCA
1	16	-5.21	-3.32	-2.11
	64	-0.86	-1.78	-1.65
	128	-0.69	-0.82	0.12
	512	0.01	0.09	0.15
2	16	0.1	0.12	0.32
	64	0.07	0.15	0.29
	128	-1.24	-0.32	0.22
	512	-0.56	-0.14	-0.05
3	16	-1.56	-0.08	0.05
	64	0.05	0.18	0.15
	128	-0.16	-0.24	-0.21
	512	-1.14	-6.69	-5.62

Table S8. Validation of the performances for the human health model, developed with different model settings and full descriptors, feature selected descriptors and PCA descriptors as input.

Number of Hidden Layer(s)	Number of Hidden Neuron (s)	Full Descriptors	Feature Selection	PCA
1	16	0.15	0.04	0.16
	64	0.12	0.14	0.18
	128	0.18	0.26	0.22
	512	0.29	0.32	0.35
2	16	0.15	0.22	0.15

	64	0.4	0.12	0.11
	128	0.4	0.29	0.35
	512	0.46	0.25	0.33
3	16	0.16	0.28	0.22
	64	0.42	0.15	0.52
	128	0.35	0.13	0.52
	512	0.26	0.08	0.33

Table S9. Validation of the performances for the ecosystem quality model, developed with different model settings and full descriptors, feature selected descriptors and PCA descriptors as input.

Number of Hidden Layer(s)	Number of Hidden Neuron (s)	Full Descriptors	Feature Selection	PCA
1	16	-0.13	0.05	-0.33
	64	0.16	0.25	-0.29
	128	0.11	0.15	0.10
	512	0.05	0.12	0.18
2	16	0.25	0.15	-0.05
	64	0.18	0.31	0.29
	128	0.21	0.26	0.35
	512	-0.04	0.08	0.05
3	16	0.14	0.18	0.15
	64	0.28	0.32	0.32
	128	0.26	0.31	0.22
	512	0.11	0.39	-0.71

According to the result in Table S4 – S9, we selected the model setting for each impact category that exhibits the highest R^2 values. The parameters we used in this study to develop the six ANNs models are presented in Table S10, including the model structure, activation function, learning rate, learning epoch and regularization factor during training.

Table S10. The hyper-parameters applied to develop ANNs models for each impact category.

	Number of Hidden Layer	Number of Hidden Neuron	Activation Function*	Learning Rate	Learning Epoch	Regularization Factor
CED	one	128	relu	0.01	500	0.01
Acidification	two	128	sigmoid	0.01	500	0.01
global warming	two	16	relu	0.001	800	0.01
EI99	two	64	sigmoid	0.01	500	0.01
Human Health	three	128	sigmoid	0.001	500	0.01
Ecosystem Quality	two	128	sigmoid	0.001	500	0.01

* The activation function is applied to every hidden layer.

Model Applicability Domain Measurement Results. The idea of using Euclidean distance as a matric for AD measurement is presented in Table S2. The applicability domain (AD) measurements for each of the six models are presented in Table S11 to Table S16.

In each table, the MRE values of the chemicals in validation dataset are reported for each impact category. The MREs are reported in two parts: the chemical within and outside the corresponding AD. This is determined by comparing the distance to the training data centroid and the selected cut-off thresholds. If, for one testing chemical, the

distance to the training data centroid is smaller than corresponding cut-off threshold, this chemical is considered within the AD.

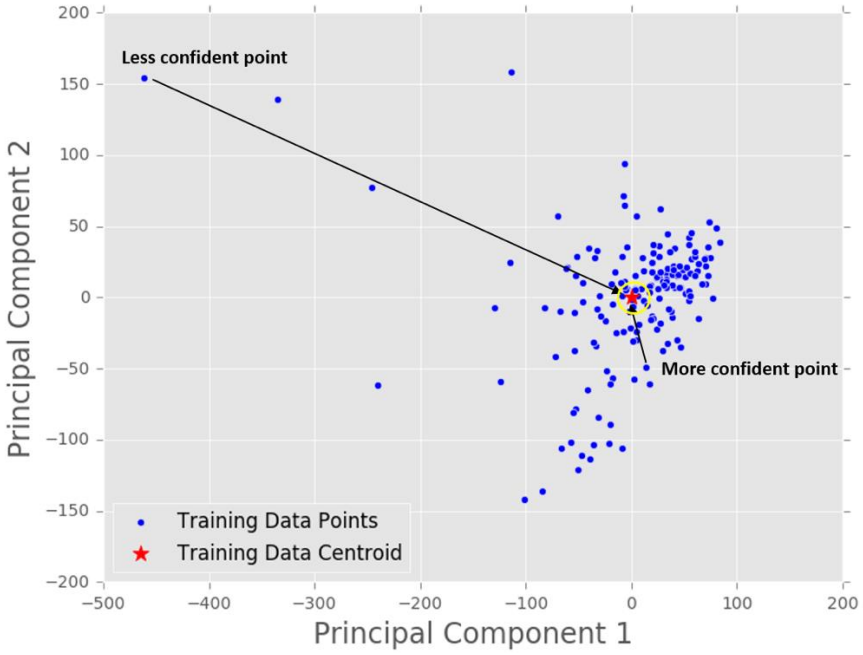

Figure S3. Projection of the collected chemical descriptors onto two-dimensional spaces by principal component analysis (PCA). The red star-point is the training data centroid. This figure illustrates the idea of distance-based AD measurement. Query chemicals that are closer to the training data centroid are more likely to have more accurate estimates than chemicals that are far away from the training data.

Table S11. Model AD measurement for the CED model with different cut-off thresholds on validation dataset.

Cut-off Threshold	MRE of Chemical within AD	Number of Chemical within AD	MRE of Chemical outside AD	Number of Chemical outside AD
500	23.9%	1	40.9%	15
600	18.7%	2	42.9%	14
700	18.7%	2	42.9%	14
800	19.2%	3	44.6%	13
900	17.8%	4	47.2%	12
1000	17.8%	4	47.2%	12
1100	25.2%	5	46.5%	11
1200	25.2%	5	46.5%	11
1300	41.3%	7	38.8%	9
1400	40.3%	8	39.4%	8
1500	52.0%	10	19.7%	6

Table S12. Model AD measurement for the acidification model with different cut-off thresholds validation dataset.

Cut-off Threshold	MRE of Chemical within AD	Number of Chemical within AD	MRE of Chemical outside AD	Number of Chemical outside AD
500	23.9%	1	40.9%	15
600	18.7%	2	42.9%	14
700	18.7%	2	42.9%	14
800	19.2%	3	44.6%	13
900	17.8%	4	47.2%	12
1000	17.8%	4	47.2%	12
1100	25.2%	5	46.5%	11
1200	25.2%	5	46.5%	11
1300	41.3%	7	38.8%	9
1400	40.3%	8	39.4%	8
1500	52.0%	10	19.7%	6

Table S13. Model AD measurement for the EI99 model with different cut-off thresholds validation dataset.

Cut-off Threshold	MRE of Chemical within AD	Number of Chemical	MRE of Chemical outside AD	Number of Chemical
500	17.7%	1	52.3%	15
600	17.7%	1	52.3%	15
700	52.8%	4	49.2%	12
800	47.3%	6	51.8%	10
900	42.5%	8	57.6%	8
1000	42.5%	8	57.6%	8
1100	39.1%	9	64.3%	7
1200	39.1%	9	64.3%	7

1300	42.2%	11	64.4%	5
1400	36.8%	13	107.8%	3
1500	36.8%	13	107.8%	3

Table S14. Model AD measurement for the global warming model with different cut-off thresholds validation dataset.

Cut-off Threshold	MRE of Chemical within AD	Number of Chemical	MRE of Chemical outside AD	Number of Chemical
500	20.5%	1	92.6%	15
600	25.1%	2	92.1%	14
700	25.1%	2	92.1%	14
800	200.0%	3	62.3%	13
900	158.6%	4	64.7%	12
1000	158.6%	4	64.7%	12
1100	144.9%	5	62.3%	11
1200	144.9%	5	62.3%	11
1300	133.2%	7	53.1%	9
1400	122.1%	8	54.2%	8
1500	119.0%	10	36.8%	6

Table S15. Model AD measurement for the human health model with different cut-off thresholds validation dataset.

Cut-off Threshold	MRE of Chemical within AD	Number of Chemical	MRE of Chemical outside AD	Number of Chemical
500	7.2%	1	130.7%	15
600	31.1%	2	136.2%	14
700	31.1%	2	136.2%	14
800	23.7%	3	145.9%	13
900	41.3%	4	150.2%	12
1000	41.3%	4	150.2%	12
1100	80.3%	5	142.4%	11
1200	80.3%	5	142.4%	11
1300	68.1%	7	165.7%	9
1400	62.8%	8	183.2%	8
1500	89.8%	10	178.4%	6

Table S16. Model AD measurement for the ecosystem quality model with different cut-off thresholds validation dataset.

Cut-off Threshold	MRE of Chemical within AD	Number of Chemical	MRE of Chemical outside AD	Number of Chemical

500	58.3%	1	52.4%	15
600	58.3%	1	52.4%	15
700	73.1%	4	46.0%	12
800	58.8%	6	49.2%	10
900	45.4%	8	60.2%	8
1000	45.4%	8	60.2%	8
1100	43.4%	9	64.9%	7
1200	43.4%	9	64.9%	7
1300	41.1%	11	78.6%	5
1400	41.0%	13	104.0%	3
1500	41.0%	13	104.0%	3

Appendix II: Supporting Information for Chapter III

Experimental Data Collection Procedure. Experimental ecotoxicity data (LC50) of organic chemicals on 8 aquatic species was collected from major public databases, such as ECOTOX, eChem, EFSA and HSDB.[114–118] Data from peer-reviewed literatures was also added as supporting data to develop the neural network models in this study.[107,110,111,119–122,183–190] The number of organic chemicals collected for 8 different species (in three taxa) is presented in Figure S1, along with the taxa information for these species.

To ensure data quality, the critical experimental conditions, such as testing duration, chemical purity and *pH* values were strictly controlled during the process of data collection. 96 hours LC50 data was used for all species except water fleas (48 hours' data was used). Chemical purity must be higher than 85%. And the *pH* value must be in the range of 5 to 9. Experimental data that not meet these requirements was discarded. For chemical with multiple experimental values, the geometric mean was used in the final dataset. The species selected in this study is aiming to cover as many aquatic taxa as possible but also should have enough experimental ecotoxicity data. After the data collection and selection, species with less than 100 unique organic chemicals' experimental values were discarded. However, to utilize some of the discarded data, experimental values that met our data selection procedure for other water fleas (*Ceriodaphnia Dubia*, *Daphnia Pulex and Mix Water Flea*) in ECOTOX database was combined and treated as an individual species in this study.

Additional information, such as the CAS number, SMILEs, molecular weight and the chemical names were also collected. The unit of the LC50 values were converted to *log10(LC50)* in *μmol/L*. The final dataset is available in the supporting information.

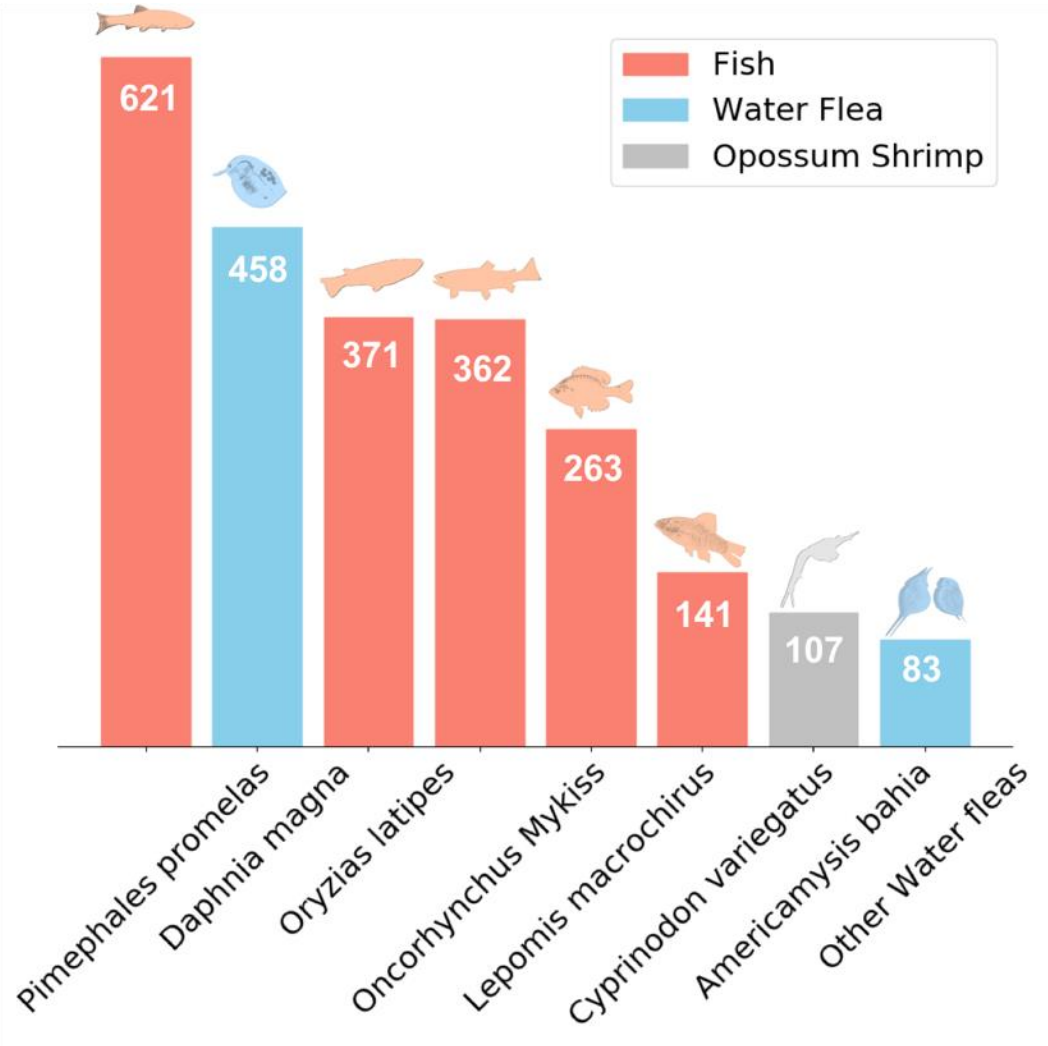

Figure S4. The number of unique chemicals collected for this study for 8 different species.

Model Performances and Hyperparameters

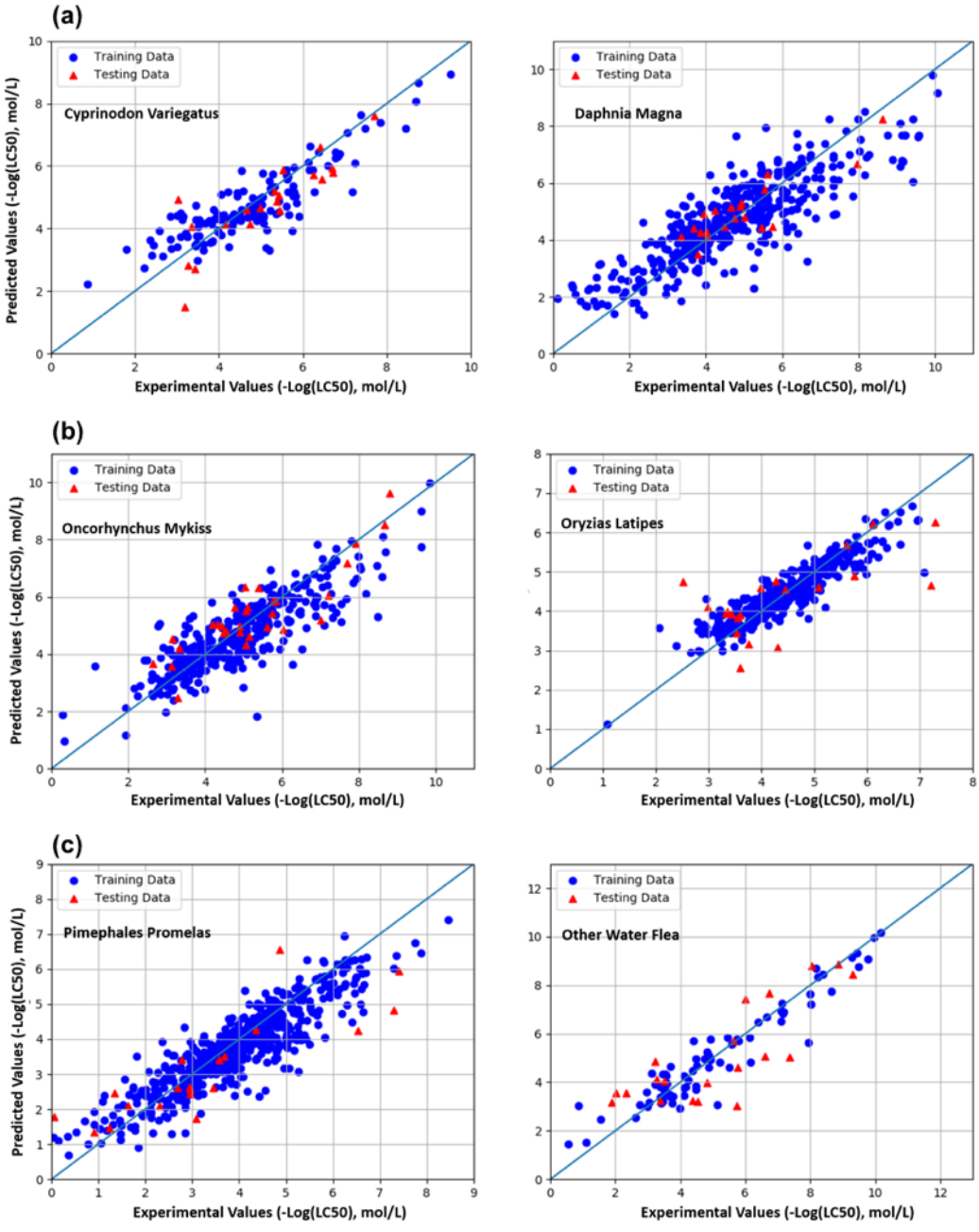

Figure S5. The performance of the Cyprinodon Variegatus and Daphnia Magna (a), Oncorhynchus Mykiss and Oryzias Latipes (b), Pimephales Promelas and Other Water Flea (c) models on the training data (blue circles) and testing data (red triangles).

Table S17. The performances (in R^2) of the QSARs on testing dataset (20 randomly selected chemicals) along with the hyper-parameters optimized in this study. For all QSARs, Rectified Linear unit (ReLu) activation function was used in hidden neuron. Learning rate was set to 0.001. The number of training iteration was 500 times.

QSAR for Species	*Pimephales Promelas*	*Daphnia Magna*	*Oryzias Latipes*	*Oncorhynchus Mykiss*	*Lepomis Macrochirus*	*Cyprinodon Variegatus*	*Americamysis Bahia*	Other Water Fleas
Model Performance (R^2) on Testing Data	0.71	0.75	0.54	0.75	0.72	0.66	0.67	0.63
Number of Hidden Layer	2	1	2	2	2	2	1	2
Number of Hidden Neuron in Each Layer	32 × 16	16	64 × 32	64 × 32	32 × 16	16 × 8	16	16 × 8
Activation Functions	ReLu, ReLu	ReLu	ReLu, ReLu	ReLu, ReLu	ReLu, ReLu	ReLu, ReLu	ReLu	ReLu, ReLu
Regularization Factor	0.01	0.02	0.01	0.02	0.03	0.05	0.01	0.05

Screening the ToX21 Database

Table S18. The top 10 chemicals among the chemicals in the ToX21 database with the lowest HC5 values according to the predictive SSDs.

Chemical Name	CAS Number	HC5 values (log(umol/L)
Dihydrostreptomycin sulfate	5490-27-7	-38.6184
Streptomycin sulfate (2:3)	3810-74-0	-37.9823
Netilmicin sulfate	56391-57-2	-36.2539
Sisomicin sulfate	53179-09-2	-33.8234
Sucrose octasulfate-aluminum complex	54182-58-0	-25.4343
Triptorelin pamoate	124508-66-3	-23.3853
YM218		-21.9758
Ergotamine D-tartrate	379-79-3	-20.7683
Pyrvinium pamoate	3546-41-6	-19.7536
Auranofin	34031-32-8	-18.8429

Model Applicable Domains

Table S19. The results of model AD analysis for each QSAR in this study. The cut-off threshold determines whether chemicals fall inside or outside a model's AD base on its distance to the centriole of the training data. The average mean square errors (MSEs) of the chemicals in the testing data that are inside and outside the AD of each model are also reported in the table.

QSAR for Species	AD Cut-off Threshold (K)	Average MSE Inside AD	Average MSE Outside AD
Pimephales Promelas	3	8%	220%
Daphnia Magna	2.5	7%	12%
Oryzias Latipes	1.5	8%	19%
Oncorhynchus Mykiss	1	6%	15%
Lepomis Macrochirus	1	6%	22%
Cyprinodon Variegatus	2.5	7%	16%
Americamysis Bahia	2	17%	19%
Other Water Fleas	3	22%	32%

Table S20. The results of model AD analysis for each QSAR in this study. The cut-off threshold determines whether chemicals fall inside or outside a model's AD base on its distance to the centriole of the training data. The average mean square errors (MSEs) of the chemicals in the testing data that are inside and outside the AD of each model are also reported in the table.

QSAR for Species	AD Cut-off Threshold (K)	Average MSE Inside AD	Average MSE Outside AD
Pimephales Promelas	3	8%	220%
Daphnia Magna	2.5	7%	12%
Oryzias Latipes	1.5	8%	19%
Oncorhynchus Mykiss	1	6%	15%
Lepomis Macrochirus	1	6%	22%
Cyprinodon Variegatus	2.5	7%	16%
Americamysis Bahia	2	17%	19%
Other Water Fleas	3	22%	32%

Comparing Predictive SSDs with Experimental SSDs

Table S21. The predictions of the ANN models for the 10 selected chemicals. The unit is log(μmol/L)

	Clofenotane	**Pentachlorophenol**	**Lindane**	**Propiconazole**	**Imidacloprid**	**Endosulfan**	**Chlorpyrifos**	**Fluoranthene**	**Aniline**	**Diazinon**
	50-29-3	**87-86-5**	**58-89-9**	**60207-90-1**	**138261-41-3**	**115-29-7**	**2921-88-2**	**206-44-0**	**62-53-3**	**333-41-5**
Americamysis Bahia	-2.8529	-0.2589	-0.6759	-0.2245	0.8985	-2.0640	-1.3199	-0.9427	2.0983	-2.2168
Lepomis Macrochirus	-1.2054	-0.0014	0.2492	-0.0044	1.6714	-1.6125	-0.6156	0.4946	2.1022	-0.9079
Oncorhynchus Mykiss	-1.1660	0.0241	0.4198	0.1999	1.7069	-1.1038	-0.5483	0.7428	2.3500	-0.7943
Cyprinodon Variegatus	-0.7202	0.2318	0.8196	0.2381	1.7209	-0.3951	-0.3919	0.7745	2.4209	0.5804
Daphnia Magna	-0.6343	0.3607	0.9857	0.4997	1.8418	0.1452	-0.1502	0.9271	2.4288	0.6645
Pimephales Promelas	0.0526	0.4146	1.7567	0.5351	1.8915	0.2450	-0.1034	1.0595	2.4303	0.8129
Oryzias Latipes	0.3035	0.6400	2.4230	1.3133	2.5427	0.5154	0.8825	1.0844	2.5558	0.8664
Other water fleas	2.6481	1.1353	4.2927	2.5075	4.4979	0.6675	2.0752	2.9870	3.4299	1.7595

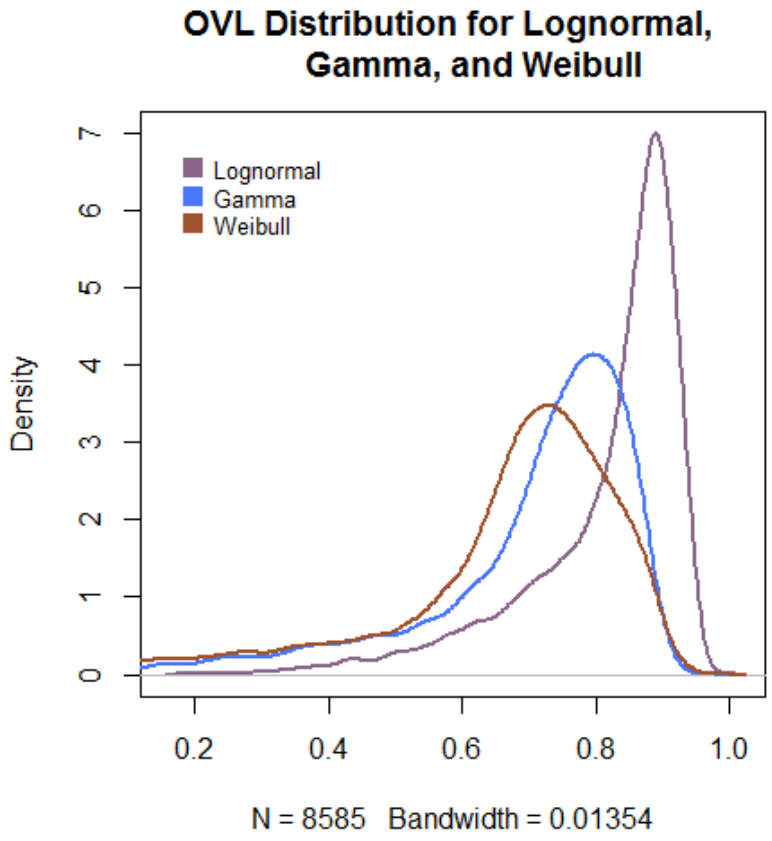

Figure S6. Comparison between log-normal, Weibull and Gamma distributions in OVL testing for the SSDs of the ToX21 chemicals.

Table S22. OVL scores for Log-normal, Gamma and Weibull distributions.

Average OVL		
Lognormal	Gamma	Weibull
81.7%	70.8%	67.2%

The Descriptors Used to Develop ANN Models for Each Species

Table S23. The full list of descriptors used to develop each ANN model.

Pimephales Promelas	Daphnia Magna	Oryzias Latipes	Oncorhynchus Mykiss	Lepomis Macrochirus	Cyprinodon Variegatus	Americamysis Bahia	Other Water Fleas
SLogP	SLogP	SLogP	SLogP	SLogP	SLogP	SLogP	SLogP
Xp-2dv	Xp-2dv	Xp-2dv	AATS3i	ATS1m	SMR_VSA4	PEOE_VSA6	MWC03
PEOE_VSA6	SM1_Dzm	PEOE_VSA6	ATS0m	ATS2m	SM1_Dzm	SMR_VSA4	AATS5v
Sm	Xp-4dv	Sm	ATSC2p	ATS3m	Xp-4dv	AATS8i	AATS6i
AATS0i	MWC03	AATS3v	ETA_eta	ATS4m	ATS1m	ATSC1m	ATSC1dv
ATS0p	AATS6v	BertzCT	MWC05	C3SP3	ATSC2m	ATSC3v	ATSC1m
ATS3m	ATS2m	nCl	SlogP_VSA4	NssO	ATSC4dv	ATSC6m	ATSC5dv
ATS5m	ATS3m	PEOE_VSA1	Xc-3d	Xp-3dv	ETA_eta	C3SP2	ETA_alpha
piPC3	ATS5m	WPath		Xp-5dv	GGI1	nRot	MAXdO
VR3_Dzi	BCUTv-1h	Xp-0dv			MID_N	SlogP_VSA11	piPC7

ESta te_VSA8	ZMI C2		PEOE_V SA3	SMR_ VSA9	Xpc-4dv ZMIC2
IC4			SpAbs_ D	Xpc-4dv	ZMIC 4
MPC 5			SRW05	ZMIC2	
Slog P_VSA11			Xpc-4d		
Zagr eb1			Xpc-4dv		

Table S24. The full name of the descriptors.

Abbreviation	Full Name
SLogP	Wildman-Crippen LogP
Xp-2dv	2-ordered Chi path weighted by valence electrons
PEOE_VSA6	MOE Charge VSA Descriptor 6 (-0.10 <= x < -0.05)
Sm	sum of constitutional weighted by mass
AATS0i	averaged moreau-broto autocorrelation of lag 0 weighted by ionization potential
ATS0p	moreau-broto autocorrelation of lag 0 weighted by polarizability
ATS3m	moreau-broto autocorrelation of lag 3 weighted by mass
ATS5m	moreau-broto autocorrelation of lag 5 weighted by mass
piPC3	3-ordered pi-path count (log scale)
VR3_Dzi	logarithmic Randic-like eigenvector-based index from Barysz matrix weighted by ionization potential
ZMIC2	2-ordered Z-modified information content
SM1_Dzm	spectral moment from Barysz matrix weighted by mass
Xp-4dv	4-ordered Chi path weighted by valence electrons

MWC03	walk count (leg-3)
AATS6v	averaged moreau-broto autocorrelation of lag 6 weighted by vdw volume
ATS2m	moreau-broto autocorrelation of lag 2 weighted by mass
BCUTv-1h	first heighest eigenvalue of Burden matrix weighted by vdw volume
EState_VSA8	EState VSA Descriptor 8 (2.05 <= x < 4.69)
IC4	4-ordered neighborhood information content
MPC5	5-ordered path count
SlogP_VSA11	MOE logP VSA Descriptor 11 (0.50 <= x < 0.60)
Zagreb1	Zagreb index (version 1)
AATS3v	averaged moreau-broto autocorrelation of lag 3 weighted by vdw volume
BertzCT	Bertz CT
nCl	number of Cl atoms
PEOE_VSA1	MOE Charge VSA Descriptor 1 (-inf < x < -0.30)
WPath	Wiener index
Xp-0dv	0-ordered Chi path weighted by valence electrons
AATS3i	averaged moreau-broto autocorrelation of lag 3 weighted by ionization potential
ATS0m	moreau-broto autocorrelation of lag 0 weighted by mass
ATSC2p	centered moreau-broto autocorrelation of lag 2 weighted by polarizability
ETA_eta	ETA composite index for reference graph
MWC05	walk count (leg-5)
SlogP_VSA4	MOE logP VSA Descriptor 4 (0.00 <= x < 0.10)
Xc-3d	3-ordered Chi cluster weighted by sigma electrons
ATS1m	moreau-broto autocorrelation of lag 1 weighted by mass
ATS4m	moreau-broto autocorrelation of lag 4 weighted by mass
C3SP3	SP3 carbon bound to 3 other carbons
NssO	number of ssO
Xp-3dv	3-ordered Chi path weighted by valence electrons

Xp-5dv	5-ordered Chi path weighted by valence electrons
SMR_VSA4	MOE MR VSA Descriptor 4 (2.24 <= x < 2.45)
ATSC2m	centered moreau-broto autocorrelation of lag 2 weighted by mass
ATSC4dv	centered moreau-broto autocorrelation of lag 4 weighted by valence electrons
GGI1	1-ordered raw topological charge
MID_N	molecular ID on N atoms
PEOE_VSA3	MOE Charge VSA Descriptor 3 (-0.25 <= x < -0.20)
SpAbs_D	graph energy from distance matrix
SRW05	walk count (leg-5, only self returning walk)
Xpc-4d	4-ordered Chi path-cluster weighted by sigma electrons
Xpc-4dv	4-ordered Chi path-cluster weighted by valence electrons
AATS8i	averaged moreau-broto autocorrelation of lag 8 weighted by ionization potential
ATSC1m	centered moreau-broto autocorrelation of lag 1 weighted by mass
ATSC3v	centered moreau-broto autocorrelation of lag 3 weighted by vdw volume
ATSC6m	centered moreau-broto autocorrelation of lag 6 weighted by mass
C3SP2	SP2 carbon bound to 3 other carbons
nRot	rotatable bonds count
SMR_VSA9	MOE MR VSA Descriptor 9 (3.80 <= x < 4.00)
AATS5v	averaged moreau-broto autocorrelation of lag 5 weighted by vdw volume
AATS6i	averaged moreau-broto autocorrelation of lag 6 weighted by ionization potential
ATSC1dv	centered moreau-broto autocorrelation of lag 1 weighted by valence electrons
ATSC5dv	centered moreau-broto autocorrelation of lag 5 weighted by valence electrons
ETA_alpha	ETA core count

MAXdO	max of dO
piPC7	7-ordered pi-path count (log scale)
ZMIC4	4-ordered Z-modified information content

Appendix III: Supporting Information for Chapter IV

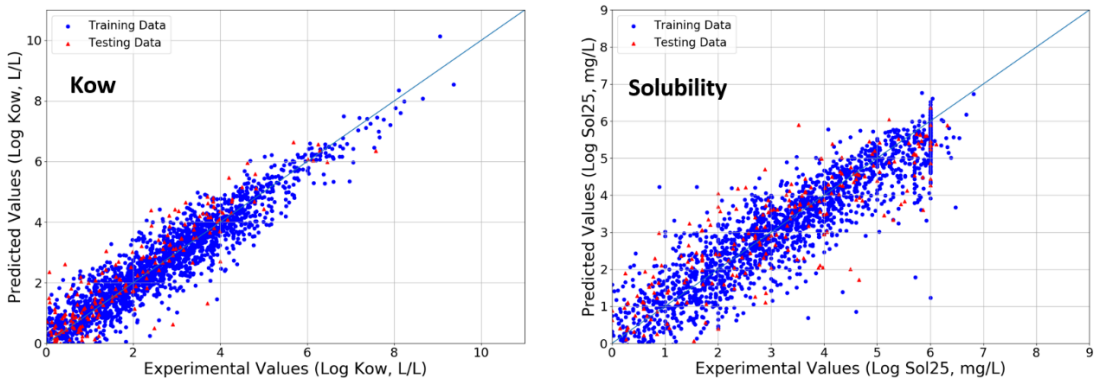

Figure S7. Model Performances for Kow (left) and Solubility (right). Both models are in ANN.

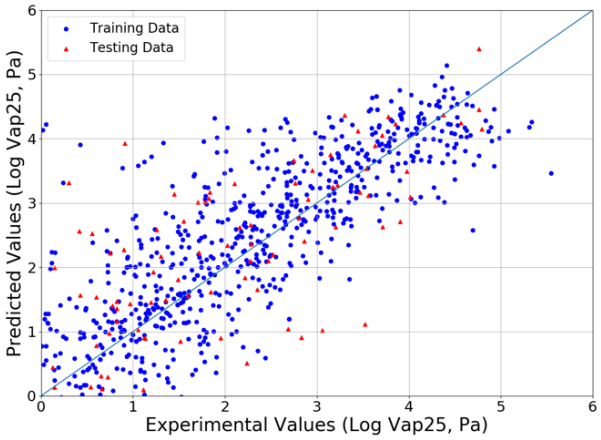

Figure S8. Model performance for Vapor Pressure, developed in ANN.

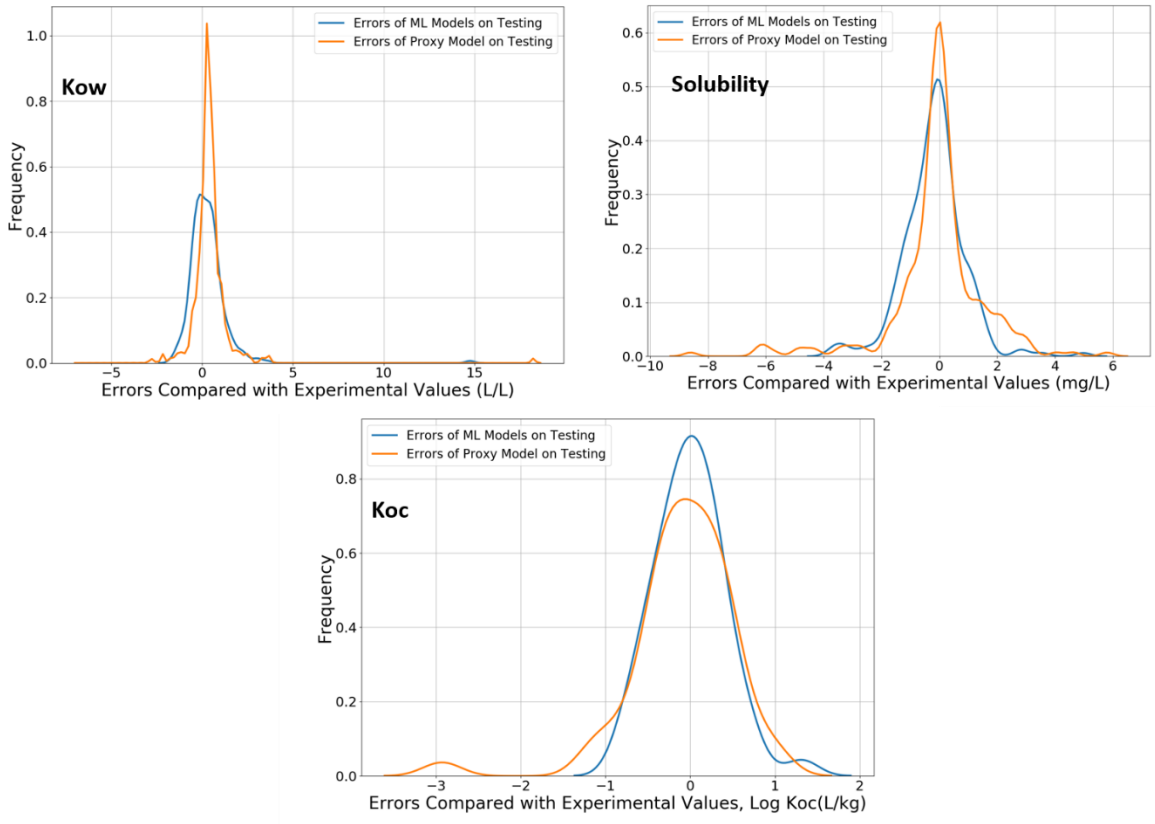

Figure S9. The absolute errors between the default proxy methods and the experimental value (orange), and between the machine learning proxy methods and the experimental values (blue), for K_{ow}, Solubility and K_{oc}.

Table S25. The FFs, EFs, XFs and CFs to freshwater compartment in North America of 383 organic chemicals that falls into the applicable domain of the ANN model in the second chapter.

CAS	Name	SMILEs	FF	EFs	XF	CFs
38083-17-9	Climbazole	CC(C)(C)C(=O)C(OC1=CC=C(Cl)C=C1)N1C=CN=C1	137.399507	6025.62536	1	8.28E+05
1713-15-1	2,4-D-isobutyl	CC(C)COC(=O)COC1=C(Cl)C=C(Cl)C=C1	50.1841884	6022.38941	1	3.02E+05
546-71-4	Ethyl 4-nitrophenyl ethylphosphonate	CCOP(=O)(CC)OC1=CC=C(C=C1)[N+]([O-])=O	49.2234211	4852.25029	1	2.39E+05
21245-02-3	2-Ethylhexyl 4-(dimethylamino)benzoate	CCCCC(CC)COC(=O)C1=CC=C(C=C1)N(C)C	94.0167286	4588.17147	1	4.31E+05
58-54-8	Ethacrynic acid	CCC(=C)C(=O)C1=CC=C(OCC(O)=O)C(Cl)=C1Cl	75.2435117	4431.34217	1	3.33E+05
3736-81-0	Diloxanide furoate	CN(C(=O)C(Cl)Cl)C1=CC=C(OC(=O)C2=CC=CO2)C=C1	124.088896	4238.51326	1	5.26E+05
5153-25-3	2-Ethylhexylparaben	CCCCC(CC)COC(=O)C1=CC=C(O)C=C1	52.8958665	2641.14846	1	1.40E+05
43076-61-5	4'-Tert-butyl-4-chlorobutyrophenone	CC(C)(C)C1=CC=C(C=C1)C(=O)CCCCl	37.494024	2471.35145	1	9.27E+04
305-03-3	Chlorambucil	OC(=O)CCCC1=CC=C(C=C1)N(CCCl)CCCl	82.0115076	2404.02025	1	1.97E+05
71868-10-5	2-Methyl-4'-(methylthio)-2-morpholinopropiophenone	CSC1=CC=C(C=C1)C(=O)C(C)(C)N1CCOCC1	156.286271	2322.20739	1	3.63E+05
149-16-6	Butacaine	CCCCN(CCCC)CCCOC(=O)C1=CC=C(N)C=C1	121.288526	2118.97983	1	2.57E+05
255714-11-5	3,7-Dimethyloct-6-en-1-yl 2-methylbut-2-enoate	CC=C(C)C(=O)OCCC(C)CCC=C(C)C	38.1380705	2108.22366	1	8.04E+04
519-88-0	Ambucetamide	CCCCN(CCCC)C(C(N)=O)C1=CC=C(OC)C=C1	188.496502	2087.9013	1	3.94E+05
14261-75-7	Cloforex	CCOC(=O)NC(C)(C)CC1=CC=C(Cl)C=C1	50.348636	1992.31616	1	1.00E+05
28730-17-8	Methfuroxam	CC1=C(C)C(C(=O)NC2=CC=CC=C2)=C(C)O1	84.8221489	1984.14223	1	1.68E+05
118-60-5	2-Ethylhexyl salicylate	CCCCC(CC)COC(=O)C1=C(O)C=CC=C1	46.8836791	1946.57252	1	9.13E+04
61570-90-9	Tioxidazole	CCCOC1=CC=C2N=C(NC(=O)OC)SC2=C1	107.551118	1907.17903	1	2.05E+05
1577-03-3	1-(4-Chlorophenyl)-4,4-dimethylpent-1-en-3-one	CC(C)(C)C(=O)C=CC1=CC=C(Cl)C=C1	53.9030296	1835.25888	1	9.89E+04
1939-27-1	3-Trifluoromethylisobutyranilide	CC(C)C(=O)NC1=CC=CC(=C1)C(F)(F)F	39.0221403	1826.38553	1	7.13E+04
13114-72-2	N'-Methyl-N,N-diphenylurea	CNC(=O)N(C1=CC=CC=C1)C1=CC=CC=C1	75.9645054	1824.1212	1	1.39E+05
40828-46-4	Suprofen	CC(C(O)=O)C1=CC=C(C=C1)C(=O)C1=CC=CS1	96.5930882	1815.62735	1	1.75E+05
61295-41-8	3-(2-Methyl-3-furylthio)-4-heptanone	CCCC(=O)C(CC)SC1=C(C)OC=C1	42.9928367	1812.48833	1	7.79E+04
71617-10-2	Amiloxate	COC1=CC=C(C=CC(=O)OCCC(C)C)C=C1	50.0293843	1801.60554	1	9.01E+04
2493-84-7	4-Octyloxybenzoic acid	CCCCCCCCOC1=CC=C(C=C1)C(O)=O	70.9739989	1794.31599	1	1.27E+05
97-32-5	4-Methoxy-3-nitro-N-phenylbenzamide	COC1=CC=C(C=C1[N+]([O-])=O)C(=O)NC1=CC=CC=C1	102.350299	1762.78392	1	1.80E+05
7785-33-3	Geranyl tiglate	C\C=C(/C)C(=O)OC\C=C(/C)CCC=C(C)C	25.5225972	1724.94988	1	4.40E+04

1219-38-1	Octylparaben	CCCCCCCCOC(=O)C1=CC=C(O)C=C1	78.9382755	1687.94407	1	1.33E+05
96568-04-6	Ethyl 4-(2,6-dichloro-5-fluoropyridin-3-yl)-3-oxobutanoate	CCOC(=O)CC(=O)C1=CC(F)=C(Cl)N=C1Cl	25.6195252	1671.49706	1	4.28E+04
35256-85-0	Butam	CC(C)N(CC1=CC=CC=C1)C(=O)C(C)(C)C	35.3552304	1624.05526	1	5.74E+04
98730-04-2	Benoxacor	CC1COC2=C(C=CC=C2)N1C(=O)C(Cl)Cl	38.9531925	1620.55227	1	6.31E+04
4252-78-2	2,2',4'-Trichloroacetophenone	ClCC(=O)C1=C(Cl)C=C(Cl)C=C1	14.7650661	1606.82363	1	2.37E+04
21440-97-1	Brofoxine	CC1(C)OC(=O)NC2=C1C=C(Br)C=C2	35.4187448	1558.85648	1	5.52E+04
66346-01-8	1-(4-Chlorophenyl)-4,4-dimethyl-pentan-3-one	CC(C)(C)C(=O)CCC1=CC=C(Cl)C=C1	41.0548447	1547.69066	1	6.35E+04
1222-98-6	4-Nitrochalcone	[O-][N+](=O)C1=CC=C(\C=C\C(=O)C2=CC=CC=C2)C=C1	91.71004	1510.6758	1	1.39E+05
784-38-3	2-Amino-5-chloro-2'-fluorobenzophenone	NC1=C(C=C(Cl)C=C1)C(=O)C1=C(F)C=CC=C1	71.8439113	1500.3768	1	1.08E+05
85-19-8	(5-Chloro-2-hydroxyphenyl)phenylmethanone	OC1=CC=C(Cl)C=C1C(=O)C1=CC=CC=C1	86.0563565	1477.21833	1	1.27E+05
602-38-0	1,8-Dinitronaphthalene	[O-][N+](=O)C1=CC=CC2=CC=CC(=C12)[N+]([O-])=O	78.9099145	1431.88414	1	1.13E+05
177785-47-6	PharmaGSID_47261	CC[C@H](C)[C@H](N1SC2=CC=CC=C2C1=O)C(O)=O	66.5246931	1416.465	1	9.42E+04
14007-64-8	Butetamate	CCC(C(=O)OCCN(CC)CC)C1=CC=CC=C1	60.7338462	1382.54397	1	8.40E+04
54965-21-8	Albendazole	CCCSC1=CC=C2NC(NC(=O)OC)=NC2=C1	101.575464	1361.38755	1	1.38E+05
637-07-0	Clofibrate	CCOC(=O)C(C)(C)OC1=CC=C(Cl)C=C1	27.7784633	1346.80887	1	3.74E+04
25059-80-7	Benazolin-ethyl	CCOC(=O)CN1C(=O)SC2=CC=CC(Cl)=C12	88.1847455	1342.63035	1	1.18E+05
101973-77-7	Esonarimod	CC(=O)SCC(CC(=O)C1=CC=C(C)C=C1)C(O)=O	63.9651382	1315.93053	1	8.42E+04
35948-25-5	6H-Dibenzo[c,e][1,2]oxaphosphinine 6-oxide	O=P1OC2=C(C=CC=C2)C2=C1C=CC=C2	82.0320845	1299.74014	1	1.07E+05
68157-60-8	Forchlorfenuron	ClC1=NC=CC(NC(=O)NC2=CC=CC=C2)=C1	81.5833649	1296.66072	1	1.06E+05
719-59-5	2-Amino-5-chlorobenzophenone	NC1=CC=C(Cl)C=C1C(=O)C1=CC=CC=C1	78.4655035	1283.37127	1	1.01E+05
15301-40-3	Actinoquinol	CCOC1=CC=C(C2=C1N=CC=C2)S(O)(=O)=O	79.580404	1278.18971	1	1.02E+05
2759-71-9	Cypromid	ClC1=CC=C(NC(=O)C2CC2)C=C1Cl	25.7110273	1236.01109	1	3.18E+04
14062-23-8	Felbinac ethyl	CCOC(=O)CC1=CC=C(C=C1)C1=CC=CC=C1	92.3622625	1226.48695	1	1.13E+05
4638-48-6	5-Chlorosalicylanilide	OC1=C(C=C(Cl)C=C1)C(=O)NC1=CC=CC=C1	72.070504	1226.01639	1	8.84E+04
606-37-1	1,3-Dinitronaphthalene	[O-][N+](=O)C1=CC2=CC=CC=C2C(=C1)[N+]([O-])=O	85.0950552	1192.45176	1	1.01E+05
605-71-0	1,5-Dinitronaphthalene	[O-][N+](=O)C1=CC=CC2=C(C=CC=C12)[N+]([O-])=O	88.5589987	1177.28388	1	1.04E+05
3575-80-2	Methylperone	CC1CCN(CCCC(=O)C2=CC=C(F)C=C2)CC1	113.161149	1154.90738	1	1.31E+05
53786-45-1	Ethyl 4-(2-amino-4-chloroanilino)piperidine-1-carboxylate	CCOC(=O)N1CCC(CC1)NC1=C(N)C=C(Cl)C=C1	158.95628	1141.00149	1	1.81E+05
69956-77-0	Pelubiprofen	CC(C(O)=O)C1=CC=C(\C=C2/CCCCC2=O)C=C1	154.293645	1116.73048	1	1.72E+05
83471-41-4	Pincainide	CC1=CC=CC(C)=C1NC(=O)CN1CCCCCC1	114.796501	1110.96104	1	1.28E+05
87-29-6	Cinnamyl anthranilate	NC1=C(C=CC=C1)C(=O)OCC=CC1=CC=CC=C1	129.877469	1110.22103	1	1.44E+05
127-63-9	Diphenylsulfone	O=S(=O)(C1=CC=CC=C1)C1=CC=CC=C1	73.7952507	1103.85557	1	8.15E+04

1085-12-7	Heptylparaben	CCCCCCCOC(=O)C1=CC=C(O)C=C1	55.666894	1065.93541	1	5.93E+04
106-29-6	Geranyl butyrate	CCCC(=O)OC\C=C(/C)CCC=C(C)C	31.9728944	1059.26416	1	3.39E+04
91374-21-9	Ropinirole	CCCN(CCC)CCC1=C2CC(=O)NC2=CC=C1	82.642163	1040.56297	1	8.60E+04
50528-97-7	Xilobam	CN1CCC=C1NC(=O)NC1=C(C)C=CC=C1C	84.4015142	1040.16703	1	8.78E+04
84-79-7	Lapachol	CC(C)=CCC1=C(O)C(=O)C2=C(C=CC=C2)C1=O	88.0928575	1028.77332	1	9.06E+04
2164-09-2	Chloranocryl	CC(=C)C(=O)NC1=CC=C(Cl)C(Cl)=C1	27.6179289	1025.1644	1	2.83E+04
10094-34-5	1,1-Dimethyl-2-phenylethyl butanoate	CCCC(=O)OC(C)(C)CC1=CC=CC=C1	26.626356	1013.92055	1	2.70E+04
957-56-2	Fluindione	FC1=CC=C(C=C1)C1C(=O)C2=CC=CC=C2C1=O	97.0691218	1012.46311	1	9.83E+04
61-68-7	Mefenamic acid	CC1=C(C)C(NC2=C(C=CC=C2)C(O)=O)=CC=C1	93.2802093	999.393445	1	9.32E+04
3562-99-0	Menbutone	COC1=CC=C(C(=O)CCC(O)=O)C2=C1C=CC=C2	148.739771	990.221419	1	1.47E+05
204005-46-9	SU-5416	CC1=CC(C)=C(N1)C=C1C(=O)NC2=CC=CC=C12	108.60092	984.362996	1	1.07E+05
16883-16-2	Pmic chloride	CC1=C(C(Cl)=O)C(=NO1)C1=CC=CC=C1	66.7937094	980.825662	1	6.55E+04
7036-58-0	Propoxate	CCCOC(=O)C1=CN=CN1C(C)C1=CC=CC=C1	118.875419	977.580537	1	1.16E+05
4433-79-8	N-(4-Chloro-2,5-dimethoxyphenyl)-3-oxobutanamide	COC1=CC(NC(=O)CC(C)=O)=C(OC)C=C1Cl	64.9741833	974.135022	1	6.33E+04
127-77-5	Sulfabenz	NC1=CC=C(C=C1)S(=O)(=O)NC1=CC=CC=C1	111.857658	973.420485	1	1.09E+05
122-40-7	Pentylcinnamaldehyde	CCCCCC(C=O)=CC1=CC=CC=C1	45.2247347	967.3453	1	4.37E+04
364-62-5	Metoclopramide	CCN(CC)CCNC(=O)C1=CC(Cl)=C(N)C=C1OC	138.118878	949.364863	1	1.31E+05
19504-77-9	Variotin	CCCCC(O)\C=C(/C)\C=C\C=C\C(=O)N1CCCC1=O	113.151116	947.817044	1	1.07E+05
1210-35-1	Dibenzosuberone	O=C1C2=CC=CC=C2CCC2=C1C=CC=C2	44.5399381	945.621774	1	4.21E+04
7779-65-9	3-Methylbutyl cinnamate	CC(C)CCOC(=O)C=CC1=CC=CC=C1	38.216451	940.154432	1	3.59E+04
21245-01-2	Padimate	CC(C)CCOC(=O)C1=CC=C(C=C1)N(C)C	45.6468657	938.498871	1	4.28E+04
42924-53-8	Nabumetone	COC1=CC2=CC=C(CCC(C)=O)C=C2C=C1	71.2247036	936.166171	1	6.67E+04
773-76-2	Chloroxine	OC1=C2N=CC=CC2=C(Cl)C=C1Cl	34.7449688	916.288189	1	3.18E+04
22494-42-4	Diflunisal	OC(=O)C1=C(O)C=CC(=C1)C1=C(F)C=C(F)C=C1	108.028718	905.590335	1	9.78E+04
60719-82-6	Alaproclate	CC(N)C(=O)OC(C)(C)CC1=CC=C(Cl)C=C1	51.1447503	903.544267	1	4.62E+04
605-45-8	Diisopropyl phthalate	CC(C)OC(=O)C1=CC=CC=C1C(=O)OC(C)C	53.9651623	882.326832	1	4.76E+04
17969-20-9	Fenclozic acid	OC(=O)CC1=CSC(=N1)C1=CC=C(Cl)C=C1	64.3165134	877.432366	1	5.64E+04
787-93-9	Ameltolide	CC1=CC=CC(C)=C1NC(=O)C1=CC=C(N)C=C1	97.7031932	874.255112	1	8.54E+04
4273-98-7	2-(Phenylsulfonyl)aniline	NC1=CC=CC=C1S(=O)(=O)C1=CC=CC=C1	86.6362641	860.062591	1	7.45E+04
94-23-5	Parethoxycaine	CCOC1=CC=C(C=C1)C(=O)OCCN(CC)CC	92.8725086	839.691774	1	7.80E+04
17737-65-4	Clonixin	CC1=C(NC2=C(C=CC=N2)C(O)=O)C=CC=C1Cl	92.828019	828.5097	1	7.69E+04
484-20-8	5-Methoxypsoralen	COC1=C2C=COC2=CC2=C1C=CC(=O)O2	71.7216322	823.489669	1	5.91E+04

33643-49-1	(+)-Ketamine	CN[C@]1(CCCCC1=O)C1=CC=CC=C1Cl	46.2951015	823.437964	1	3.81E+04
6740-88-1	Ketamine	CNC1(CCCCC1=O)C1=C(Cl)C=CC=C1	46.2925913	823.437964	1	3.81E+04
33643-46-8	(S)-Ketamine	CN[C@@]1(CCCCC1=O)C1=CC=CC=C1Cl	46.2925913	823.437964	1	3.81E+04
22204-53-1	Naproxen	COC1=CC2=CC=C(C=C2C=C1)[C@H](C)C(O)=O	75.343401	818.361444	1	6.17E+04
29876-14-0	Nicotredole	O=C(NCCC1=CNC2=CC=CC=C12)C1=CC=CN=C1	137.1352	779.440571	1	1.07E+05
94-20-2	Chlorpropamide	CCCNC(=O)NS(=O)(=O)C1=CC=C(Cl)C=C1	45.0396717	773.569482	1	3.48E+04
1223-36-5	Clofexamide	CCN(CC)CCNC(=O)COC1=CC=C(Cl)C=C1	80.3671579	767.399614	1	6.17E+04
118-57-0	Acetaminosalol	CC(=O)NC1=CC=C(OC(=O)C2=CC=CC=C2O)C=C1	133.259662	764.339504	1	1.02E+05
41653-21-8	Sulcaine	CCOC(=O)C1=CC=C(NC(=O)CN2CCCCC2)C=C1	140.875127	755.77397	1	1.06E+05
51934-41-9	Benzoicacid,4-iodo-,ethylester	CCOC(=O)C1=CC=C(I)C=C1	8.90557105	746.503014	1	6.65E+03
58473-74-8	Cinromide	CCNC(=O)\C=C\C1=CC=CC(Br)=C1	20.7314854	726.93178	1	1.51E+04
4093-31-6	Methyl 4-acetamido-5-chloro-o-anisate	COC(=O)C1=C(OC)C=C(NC(C)=O)C(Cl)=C1	52.4762476	723.06146	1	3.79E+04
31431-43-3	Cyclobendazole	COC(=O)NC1=NC2=C(N1)C=CC(=C2)C(=O)C1CC1	89.9660584	722.404368	1	6.50E+04
111406-87-2	Zileuton	CC(N(O)C(N)=O)C1=CC2=C(S1)C=CC=C2	66.0811693	718.568212	1	4.75E+04
1508-75-4	Tropicamide	CCN(CC1=CC=NC=C1)C(=O)C(CO)C1=CC=CC=C1	152.798353	716.155152	1	1.09E+05
5174-32-3	2-Acetoxy-5-nitrobenzyl chloride	CC(=O)OC1=C(CCl)C=C(C=C1)[N+]([O-])=O	27.5861439	715.435747	1	1.97E+04
7303-78-8	Imidoline	CN(C)CCN1CCN(C1=O)C1=CC(Cl)=CC=C1	66.1902279	713.622519	1	4.72E+04
607-57-8	2-Nitrofluorene	[O-][N+](=O)C1=CC2=C(C=C1)C1=C(C2)C=CC=C1	89.3298184	708.246423	1	6.33E+04
15345-89-8	Desmethoxyyangonin	COC1=CC(=O)OC(\C=C\C2=CC=CC=C2)=C1	76.0504036	694.913607	1	5.28E+04
148-82-3	Melphalan	N[C@@H](CC1=CC=C(C=C1)N(CCCl)CCCl)C(O)=O	89.357779	694.763601	1	6.21E+04
94-47-3	2-Phenylethyl benzoate	O=C(OCCC1=CC=CC=C1)C1=CC=CC=C1	67.980195	694.547089	1	4.72E+04
31224-92-7	Pifoxime	C\C(=N/O)C1=CC=C(OCC(=O)N2CCCCC2)C=C1	144.156439	692.13561	1	9.98E+04
1083-27-8	Hexylparaben	CCCCCCOC(=O)C1=CC=C(O)C=C1	54.0135784	685.9643	1	3.71E+04
71526-07-3	MON-4660	ClC(Cl)C(=O)N1CCOC11CCCCC1	47.139814	680.677635	1	3.21E+04
93-97-0	Benzoic anhydride	O=C(OC(=O)C1=CC=CC=C1)C1=CC=CC=C1	78.7364707	676.196082	1	5.32E+04
91-44-1	7-Diethylamino-4-methylcoumarin	CCN(CC)C1=CC=C2C(C)=CC(=O)OC2=C1	58.610264	674.956148	1	3.96E+04
31036-80-3	Lofexidine	CC(OC1=C(Cl)C=CC=C1Cl)C1=NCCN1	44.5689672	674.571964	1	3.01E+04
23049-93-6	Enfenamic acid	OC(=O)C1=C(NCCC2=CC=CC=C2)C=CC=C1	87.5978241	671.236426	1	5.88E+04
6290-37-5	2-Phenylethyl hexanoate	CCCCCC(=O)OCCC1=CC=CC=C1	53.7969558	655.179235	1	3.52E+04
3686-58-6	Tolycaine	CCN(CC)CC(=O)NC1=C(C)C=CC=C1C(=O)OC	92.4817403	651.688703	1	6.03E+04
53558-25-1	Pyrinuron	[O-][N+](=O)C1=CC=C(NC(=O)NCC2=CC=CN=C2)C=C1	85.427647	649.843383	1	5.55E+04
55066-56-3	4-Methylphenyl 3-methylbutanoate	CC(C)CC(=O)OC1=CC=C(C)C=C1	18.997198	646.957101	1	1.23E+04

21834-92-4	5-Methyl-2-phenyl-2-hexenal	CC(C)CC=C(C=O)C1=CC=CC=C1	23.5364627	645.174763	1	1.52E+04
122-67-8	Isobutyl 3-phenylacrylate	CC(C)COC(=O)C=CC1=CC=CC=C1	19.2551618	634.087142	1	1.22E+04
90-51-7	6-Amino-4-hydroxynaphthalene-2-sulfonic acid	NC1=CC2=C(C=C1)C=C(C=C2O)S(O)(=O)=O	74.821549	632.66783	1	4.73E+04
3766-60-7	Buturon	CC(C#C)N(C)C(=O)NC1=CC=C(Cl)C=C1	36.9175922	632.290448	1	2.33E+04
458-24-2	Fenfluramine	CCNC(C)CC1=CC(=CC=C1)C(F)(F)F	22.7172043	627.802902	1	1.43E+04
7654-03-7	Benmoxin	CC(NNC(=O)C1=CC=CC=C1)C1=CC=CC=C1	87.0607847	618.29529	1	5.38E+04
2122-70-5	Ethyl 1-naphthaleneacetate	CCOC(=O)CC1=CC=CC2=C1C=CC=C2	59.367638	617.48814	1	3.67E+04
66532-85-2	Propacetamol	CCN(CC)CC(=O)OC1=CC=C(NC(C)=O)C=C1	90.8821363	612.180412	1	5.56E+04
298-81-7	8-Methoxypsoralen	COC1=C2OC(=O)C=CC2=CC2=C1OC=C2	57.9591636	610.015059	1	3.54E+04
2882-19-1	Ethyl bromophenylacetate	CCOC(=O)C(Br)C1=CC=CC=C1	22.8714791	607.165997	1	1.39E+04
80-27-3	Terpinyl propionate	CCC(=O)OC(C)(C)C1CCC(C)=CC1	34.1336128	602.482654	1	2.06E+04
77671-31-9	Enoximone	CSC1=CC=C(C=C1)C(=O)C1=C(C)NC(=O)N1	83.4532991	599.618735	1	5.00E+04
6789-88-4	Hexyl benzoate	CCCCCCOC(=O)C1=CC=CC=C1	19.6124788	598.723381	1	1.17E+04
7761-45-7	Methodichlorophen	CC1=C(C(N)=NC(N)=N1)C1=CC(Cl)=C(Cl)C=C1	105.089149	587.032645	1	6.17E+04
25152-85-6	(3Z)-Hex-3-en-1-yl benzoate	CC\C=C/CCOC(=O)C1=CC=CC=C1	19.58808	581.129759	1	1.14E+04
24817-51-4	2-Phenylethyl 2-methylbutanoate	CCC(C)C(=O)OCCC1=CC=CC=C1	32.196507	573.931105	1	1.85E+04
103-95-7	2-Methyl-3-[4-(propan-2-yl)phenyl]propanal	CC(CC1=CC=C(C=C1)C(C)C)C=O	22.6752253	572.04894	1	1.30E+04
965-52-6	Nifuroxazide	OC1=CC=C(C=C1)C(=O)NN=CC1=CC=C(O1)[N+]([O-])=O	92.8874501	559.667117	1	5.20E+04
20559-55-1	Oxibendazole	CCCOC1=CC2=C(NC(NC(=O)OC)=N2)C=C1	121.231592	553.728492	1	6.71E+04
1609-66-1	Norfentanyl	CCC(=O)N(C1CCNCC1)C1=CC=CC=C1	50.5088193	553.377044	1	2.80E+04
133-18-6	Phenethyl anthranilate	NC1=CC=CC=C1C(=O)OCCC1=CC=CC=C1	106.091578	551.867739	1	5.85E+04
51146-56-6	Dexibuprofen	CC(C)CC1=CC=C(C=C1)[C@H](C)C(O)=O	43.197179	546.343234	1	2.36E+04
65405-77-8	(3Z)-Hex-3-en-1-yl salicylate	CC\C=C/CCOC(=O)C1=C(O)C=CC=C1	29.0071937	546.243966	1	1.58E+04
101-71-3	Diphenan	NC(=O)OC1=CC=C(CC2=CC=CC=C2)C=C1	69.5840799	544.272023	1	3.79E+04
102-16-9	Benzyl phenylacetate	O=C(CC1=CC=CC=C1)OCC1=CC=CC=C1	78.179632	538.727576	1	4.21E+04
64-77-7	Tolbutamide	CCCCNC(=O)NS(=O)(=O)C1=CC=C(C)C=C1	53.8813532	537.621203	1	2.90E+04
118-58-1	Benzyl salicylate	OC1=C(C=CC=C1)C(=O)OCC1=CC=CC=C1	82.9133476	537.130457	1	4.45E+04
138112-76-2	Agomelatine	COC1=CC2=C(CCNC(C)=O)C=CC=C2C=C1	95.9043697	530.418669	1	5.09E+04
4394-04-1	Metanixin	CC1=CC=CC(C)=C1NC1=NC=CC=C1C(O)=O	81.7087827	527.900641	1	4.31E+04
115-95-7	Linalyl acetate	CC(C)=CCCC(C)(OC(C)=O)C=C	16.6514684	526.975738	1	8.77E+03
105-95-3	1,4-Dioxacycloheptadecane-5,17-dione	O=C1CCCCCCCCCCCC(=O)OCCO1	62.554805	525.80538	1	3.29E+04
31188-99-5	4'-Piperidinylcarbonylmethoxyacetophenone	CC(=O)C1=CC=C(OCC(=O)N2CCCCC2)C=C1	124.917104	519.27309	1	6.49E+04

602-87-9	5-Nitroacenaphthene	[O-][N+](=O)C1=CC=C2CCC3=CC=CC1=C23	56.0769033	519.115826	1	2.91E+04
15165-67-0	Dichlorprop-P	C[C@@H](OC1=C(Cl)C=C(Cl)C=C1)C(O)=O	29.3898826	513.979637	1	1.51E+04
93-00-5	6-Aminonaphthalene-2-sulfonic acid	NC1=CC2=C(C=C1)C=C(C=C2)S(O)(=O)=O	99.2927747	511.993944	1	5.08E+04
6965-71-5	alpha-(2,5-Dichlorophenoxy)propionic acid	CC(OC1=C(Cl)C=CC(Cl)=C1)C(O)=O	29.3838795	511.663999	1	1.50E+04
68767-14-6	Loxoprofen	CC(C(O)=O)C1=CC=C(CC2CCCC2=O)C=C1	74.3713769	506.890163	1	3.77E+04
13898-58-3	Benzoylpas	OC(=O)C1=CC=C(NC(=O)C2=CC=CC=C2)C=C1O	113.280656	504.206367	1	5.71E+04
91-79-2	Thenyldiamine	CN(C)CCN(CC1=CSC=C1)C1=NC=CC=C1	69.3748631	503.677579	1	3.49E+04
67268-43-3	Giparmen	CC1=CC(=O)OC2=CC(OCC#C)=CC=C12	65.1021238	502.962103	1	3.27E+04
1118-39-4	2-methyl-6-methylideneoct-7-en-2-yl acetate	CC(=O)OC(C)(C)CCCC(=C)C=C	16.9017385	502.300931	1	8.49E+03
54-36-4	Metyrapone	CC(C)(C(=O)C1=CN=CC=C1)C1=CN=CC=C1	89.0154363	491.230193	1	4.37E+04
119-17-5	3-(3-Methyl-5-oxo-4,5-dihydro-1H-pyrazol-1-yl)benzenesulfonic acid	CC1=NN(C(=O)C1)C1=CC(=CC=C1)S(O)(=O)=O	59.5448	481.998078	1	2.87E+04
2210-77-7	Pyrrocaine	CC1=CC=CC(C)=C1NC(=O)CN1CCCC1	66.3090124	476.798293	1	3.16E+04
1137-42-4	4-Hydroxybenzophenone	OC1=CC=C(C=C1)C(=O)C1=CC=CC=C1	38.542097	475.529798	1	1.83E+04
94-18-8	Benzylparaben	OC1=CC=C(C=C1)C(=O)OCC1=CC=CC=C1	88.516715	475.302472	1	4.21E+04
149647-78-9	Suberoylanilide hydroxamic acid	ONC(=O)CCCCCCC(=O)NC1=CC=CC=C1	73.6463074	472.931973	1	3.48E+04
94-46-2	Isopentyl benzoate	CC(C)CCOC(=O)C1=CC=CC=C1	19.5132117	471.425022	1	9.20E+03
326-06-7	4,4,4-Trifluoro-1-phenyl-1,3-butanedione	FC(F)(F)C(=O)CC(=O)C1=CC=CC=C1	17.9027516	466.434064	1	8.35E+03
3615-24-5	Ramifenazone	CC(C)NC1=C(C)N(C)N(C1=O)C1=CC=CC=C1	56.9540443	461.237441	1	2.63E+04
483-63-6	Crotamiton	CCN(C(=O)C=CC)C1=CC=CC=C1C	38.4937932	458.97756	1	1.77E+04
53786-28-0	5-Chloro-1-(4-piperidyl)-1H-benzimidazol-2(3H)-one	ClC1=CC2=C(C=C1)N(C1CCNCC1)C(=O)N2	82.5271526	452.087267	1	3.73E+04
87940-60-1	Eprobemide	ClC1=CC=C(C=C1)C(=O)NCCCN1CCOCC1	128.452898	451.520269	1	5.80E+04
117-79-3	2-Aminoanthraquinone	NC1=CC=C2C(=O)C3=C(C=CC=C3)C(=O)C2=C1	141.568031	449.56503	1	6.36E+04
4394-05-2	Nixylic acid	CC1=CC=CC(NC2=NC=CC=C2C(O)=O)=C1C	81.8570538	449.180061	1	3.68E+04
3874-54-2	4-Chloro-1-(4-fluorophenyl)-1-butanone	FC1=CC=C(C=C1)C(=O)CCCCl	27.3563439	447.57191	1	1.22E+04
6606-59-3	1,6-Hexanediol dimethacrylate	CC(=C)C(=O)OCCCCCCOC(=O)C(C)=C	27.7568634	434.531773	1	1.21E+04
150824-47-8	(E)-Nitenpyram	CCN(CC1=CC=C(Cl)N=C1)C(\NC)=C\[N+]([O-])=O	55.8612802	431.219094	1	2.41E+04
103-38-8	Benzyl 3-methylbutanoate	CC(C)CC(=O)OCC1=CC=CC=C1	17.7578994	427.462545	1	7.59E+03
1907-65-9	N-Butyl-p-toluenesulfonamide	CCCCNS(=O)(=O)C1=CC=C(C)C=C1	45.078704	426.816963	1	1.92E+04
479-92-5	Propyphenazone	CC(C)C1=C(C)N(C)N(C1=O)C1=CC=CC=C1	52.3150271	419.52844	1	2.19E+04
6521-29-5	Pentylparaben	CCCCCOC(=O)C1=CC=C(O)C=C1	23.1811352	414.651215	1	9.61E+03
32838-28-1	Butoctamide semisuccinate	CCCCC(CC)CNC(=O)CC(C)OC(=O)CCC(O)=O	100.842762	410.773688	1	4.14E+04
40188-45-2	3'-Acetyl-4'-hydroxybutyranilide	CCCC(=O)NC1=CC(C(C)=O)=C(O)C=C1	54.6475146	405.033797	1	2.21E+04

18127-01-0	3-(4-tert-Butylphenyl)propanal	CC(C)(C)C1=CC=C(CCC=O)C=C1	27.7991507	404.18837	1	1.12E+04
882-09-7	Clofibric acid	CC(C)(OC1=CC=C(Cl)C=C1)C(O)=O	22.2659514	399.584509	1	8.90E+03
81-84-5	1H,3H-Naphtho(1,8-cd)pyran-1,3-dione	O=C1OC(=O)C2=C3C(C=CC=C13)=CC=C2	73.6415656	394.170637	1	2.90E+04
56326-98-8	1-(4-Fluorophenyl)-4-oxocyclohexanecarbonitrile	FC1=CC=C(C=C1)C1(CCC(=O)CC1)C#N	74.194979	389.334047	1	2.89E+04
104-28-9	Cinoxate	CCOCCOC(=O)C=CC1=CC=C(OC)C=C1	56.9054548	380.896498	1	2.17E+04
22131-79-9	Alclofenac	OC(=O)CC1=CC=C(OCC=C)C(Cl)=C1	22.876604	378.407387	1	8.66E+03
4247-02-3	Isobutylparaben	CC(C)COC(=O)C1=CC=C(O)C=C1	23.8868279	373.534374	1	8.92E+03
17526-94-2	3,3'-(4-Methylbenzene-1,3-diyl)bis(1,1-dimethylurea)	CN(C)C(=O)NC1=CC=C(C)C(NC(=O)N(C)C)=C1	88.7629937	371.866796	1	3.30E+04
131-67-9	Phthalofyne	CCC(C)(OC(=O)C1=C(C=CC=C1)C(O)=O)C#C	69.7757492	367.725555	1	2.57E+04
54982-83-1	1,4-Dioxacyclohexadecane-5,16-dione	O=C1CCCCCCCCCCC(=O)OCCO1	51.8742422	365.325573	1	1.90E+04
15574-49-9	Mecarbinate	CCOC(=O)C1=C(C)N(C)C2=CC=C(O)C=C12	83.9209134	365.067599	1	3.06E+04
500-64-1	Kavain	COC1=CC(=O)O[C@H](C1)\C=C\C1=CC=CC=C1	66.6914332	357.223744	1	2.38E+04
104-27-8	1-(4-Methoxyphenyl)-1-pentene-3-one	CCC(=O)C=CC1=CC=C(OC)C=C1	22.5958772	352.383106	1	7.96E+03
587-63-3	Dihydrokavain	COC1=CC(=O)OC(CCC2=CC=CC=C2)C1	74.4156238	344.866927	1	2.57E+04
55719-85-2	Phenethyl tiglate	C\C=C(/C)C(=O)OCCC1=CC=CC=C1	20.7201504	343.622098	1	7.12E+03
120-50-3	Isobutyl benzoate	CC(C)COC(=O)C1=CC=CC=C1	15.7900234	336.908334	1	5.32E+03
2876-78-0	Methyl 1-naphthaleneacetate	COC(=O)CC1=C2C=CC=CC2=CC=C1	64.3324595	333.544688	1	2.15E+04
7011-83-8	Dihydrojasmone lactone	CCCCCCC1(C)CCC(=O)O1	24.9015664	329.468546	1	8.20E+03
947-19-3	(1-Hydroxycyclohexyl)(phenyl)methanone	OC1(CCCCC1)C(=O)C1=CC=CC=C1	39.5874318	321.185828	1	1.27E+04
122-82-7	N-(4-Ethoxyphenyl)-3-oxobutanamide	CCOC1=CC=C(NC(=O)CC(C)=O)C=C1	49.4826284	317.911735	1	1.57E+04
42482-06-4	2-Octen-1-ylsuccinic anhydride	CCCCCC=CCC1CC(=O)OC1=O	41.0648885	311.931315	1	1.28E+04
6285-05-8	Ethyl 4-chlorophenyl ketone	CCC(=O)C1=CC=C(Cl)C=C1	9.89883943	310.847559	1	3.08E+03
151-05-3	Dimethylbenzylcarbinyl acetate	CC(=O)OC(C)(C)CC1=CC=CC=C1	27.0273366	301.483296	1	8.15E+03
81-16-3	2-Amino-1-naphthalenesulfonic acid	NC1=CC=C2C=CC=CC2=C1S(O)(=O)=O	91.943088	299.464824	1	2.75E+04
17369-59-4	3-Propylidenephthalide	CC\C=C1\OC(=O)C2=CC=CC=C12	17.0788909	296.164181	1	5.06E+03
71475-35-9	Lozilurea	CCNC(=O)NCC1=CC(Cl)=CC=C1	21.323353	294.81334	1	6.29E+03
103-28-6	Benzyl 2-methylpropanoate	CC(C)C(=O)OCC1=CC=CC=C1	19.6744996	294.204618	1	5.79E+03
94-14-4	Isocaine	CC(C)COC(=O)C1=CC=C(N)C=C1	23.1048772	293.872373	1	6.79E+03
614-45-9	tert-Butyl perbenzoate	CC(C)(C)OOC(=O)C1=CC=CC=C1	17.8780683	293.646478	1	5.25E+03
97-42-7	Carvyl acetate	CC(=O)OC1CC(CC=C1C)C(C)=C	30.3006271	289.203467	1	8.76E+03
488-10-8	Jasmone	CC\C=C/CC1=C(C)CCC1=O	8.93468818	286.762076	1	2.56E+03
37526-88-8	Benzyl tiglate	C\C=C(/C)C(=O)OCC1=CC=CC=C1	19.6025923	286.432614	1	5.61E+03

90-87-9	2-Phenylpropionaldehyde dimethyl acetal	COC(OC)C(C)C1=CC=CC=C1	15.482 4572	286.22 9741	1	4.43E +03
87-19-4	Isobutyl salicylate	CC(C)COC(=O)C1=CC=CC=C1O	22.276 1371	285.43 1535	1	6.36E +03
14375-45-2	Abscisic acid	C\C(\C=C\C1(O)C(C)=CC(=O)CC1(C)C)=C\C(O)=O	127.84 1264	283.19 1257	1	3.62E +04
26049-70-7	2-Hydrazino-4-(4-nitrophenyl)thiazole	NNC1=NC(=CS1)C1=CC=C(C=C1)[N+]([O-])=O	42.701 4034	280.82 9101	1	1.20E +04
6175-45-7	2,2-Diethoxyacetophenone	CCOC(OCC)C(=O)C1=CC=CC=C1	25.276 5239	280.22 0504	1	7.08E +03
94-44-0	Benzyl nicotinate	O=C(OCC1=CC=CC=C1)C1=CN=CC=C1	39.688 4635	278.49 3108	1	1.11E +04
13361-34-7	2-Ethylhexyl cyanoacetate	CCCCC(CC)COC(=O)CC#N	16.702 1433	275.94 4824	1	4.61E +03
1553-60-2	Ibufenac	CC(C)CC1=CC=C(CC(O)=O)C=C1	44.883 8389	274.55 5704	1	1.23E +04
105-87-3	Geranyl acetate	CC(C)=CCC\C(C)=C\COC(C)=O	15.584 1117	273.40 1246	1	4.26E +03
141-12-8	cis-3,7-Dimethyl-2,6-octadien-1-yl acetate	CC(C)=CCC\C(C)=C/COC(C)=O	15.584 1117	273.40 1246	1	4.26E +03
34841-35-5	3'-Chloropropiophenone	CCC(=O)C1=CC(Cl)=CC=C1	8.8328 2443	272.60 8899	1	2.41E +03
120-45-6	1-Phenylethyl propionate	CCC(=O)OC(C)C1=CC=CC=C1	16.340 3947	270.73 7722	1	4.42E +03
2270-60-2	Methyl citronellate	COC(=O)CC(C)CCC=C(C)C	16.480 5298	269.60 8864	1	4.44E +03
1885-14-9	Phenyl carbonochloridate	ClC(=O)OC1=CC=CC=C1	9.2101 4328	266.75 8956	1	2.46E +03
73-31-4	Melatonin	COC1=CC2=C(NC=C2CCNC(C)=O)C=C1	75.403 833	255.57 3753	1	1.93E +04
81-83-4	1H-Benzo[de]isoquinoline-1,3(2H)-dione	O=C1NC(=O)C2=CC=CC3=C2C1=CC=C3	57.644 4471	255.15 4481	1	1.47E +04
23597-82-2	Hexyl nicotinate	CCCCCCOC(=O)C1=CN=CC=C1	29.332 1687	253.10 6208	1	7.42E +03
5205-11-8	3-Methyl-2-butenyl benzoate	CC(C)=CCOC(=O)C1=CC=CC=C1	17.574 1386	252.95 0389	1	4.45E +03
119515-38-7	Icaridin	CCC(C)OC(=O)N1CCCCC1CCO	44.763 6969	251.47 8869	1	1.13E +04
1083-57-4	3-Hydroxy-4-butyrophenetidide	CCOC1=CC=C(NC(=O)CC(C)O)C=C1	46.914 7605	251.38 4404	1	1.18E +04
89-33-8	Ethyl 5-oxo-1-phenyl-4,5-dihydro-1H-pyrazole-3-carboxylate	CCOC(=O)C1=NN(C(=O)C1)C1=CC=CC=C1	66.521 9396	249.68 7318	1	1.66E +04
97-36-9	N-(2,4-Dimethylphenyl)-3-oxobutanamide	CC(=O)CC(=O)NC1=CC=C(C)C=C1C	48.927 5495	246.25 9478	1	1.20E +04
93-65-2	Mecoprop	CC(OC1=C(C)C=C(Cl)C=C1)C(O)=O	24.336 8952	245.52 8926	1	5.98E +03
131-70-4	Monobutyl phthalate	CCCCOC(=O)C1=C(C=CC=C1)C(O)=O	36.325 9647	243.02 9464	1	8.83E +03
64379-93-7	Cinflumide	FC1=CC(\C=C\C(=O)NC2CC2)=CC=C1	25.583 5422	241.54 5219	1	6.18E +03
120-23-0	2-Naphthoxyacetic acid	OC(=O)COC1=CC2=C(C=CC=C2)C=C1	51.742 6658	241.03 9536	1	1.25E +04
31906-04-4	4-(4-Hydroxy-4-methylpentyl)cyclohex-3-ene-1-carbaldehyde	CC(C)(O)CCCC1=CCC(CC1)C=O	45.733 1583	240.47 5992	1	1.10E +04
501-53-1	Benzyl chloroformate	ClC(=O)OCC1=CC=CC=C1	11.875 8905	238.93 4991	1	2.84E +03
71320-77-9	Moclobemide	ClC1=CC=C(C=C1)C(=O)NCCN1CCOCC1	109.28 6128	237.82 943	1	2.60E +04
67883-79-8	(3Z)-Hex-3-en-1-yl (2E)-2-methylbut-2-enoate	CC\C=C/CCOC(=O)C(\C)=C\C	14.111 1301	234.09 3119	1	3.30E +03
94-26-8	Butylparaben	CCCCOC(=O)C1=CC=C(O)C=C1	23.181 1184	229.61 9251	1	5.32E +03
27247-96-7	2-Ethylhexyl nitrate	CCCCC(CC)CO[N+]([O-])=O	10.389 2297	228.63 5777	1	2.38E +03

1515-72-6	2-Butyl-1H-isoindole-1,3(2H)-dione	CCCCN1C(=O)C2=C(C=CC=C2)C1=O	44.1608024	218.086108	1	9.63E+03
16852-81-6	Benzoclidine	O=C(OC1CN2CCC1CC2)C1=CC=CC=C1	60.444802	217.822843	1	1.32E+04
115-99-1	Linalyl formate	CC(C)=CCCC(C)(OC=O)C=C	13.5845994	212.596457	1	2.89E+03
103-52-6	2-Phenylethyl butanoate	CCCC(=O)OCCC1=CC=CC=C1	21.925309	208.535611	1	4.57E+03
3572-06-3	4-(4-(Acetyloxy)phenyl)-2-butanone	CC(=O)CCC1=CC=C(OC(C)=O)C=C1	21.6632745	208.146294	1	4.51E+03
17696-61-6	sec-Butylparaben	CCC(C)OC(=O)C1=CC=C(O)C=C1	20.0374258	208.011301	1	4.17E+03
136-60-7	Butyl benzoate	CCCCOC(=O)C1=CC=CC=C1	16.4582973	205.62058	1	3.38E+03
101-10-0	Cloprop	CC(OC1=CC(Cl)=CC=C1)C(O)=O	24.3101713	201.771003	1	4.91E+03
774-55-0	6-Acetyl-1,2,3,4-tetrahydronaphthalene	CC(=O)C1=CC2=C(CCCC2)C=C1	24.1833314	201.340796	1	4.87E+03
4093-29-2	Methyl 4-acetamido-o-anisate	COC(=O)C1=C(OC)C=C(NC(C)=O)C=C1	52.5258278	198.161024	1	1.04E+04
92-15-9	N-(2-Methoxyphenyl)-3-oxobutanamide	COC1=C(NC(=O)CC(C)=O)C=CC=C1	23.9638134	197.500408	1	4.73E+03
NOCAS_47129	Methyl geranate	COC(=O)C=C(C)CCC=C(C)C	16.6114164	194.176044	1	3.23E+03
2628-16-2	4-Ethenylphenyl acetate	CC(=O)OC1=CC=C(C=C)C=C1	13.7676557	190.909234	1	2.63E+03
1134-47-0	Baclofen	NCC(CC(O)=O)C1=CC=C(Cl)C=C1	24.7947444	189.907197	1	4.71E+03
77-83-8	Ethyl methylphenylglycidate	CCOC(=O)C1OC1(C)C1=CC=CC=C1	42.9204662	185.857975	1	7.98E+03
939-97-9	4-tert-Butylbenzaldehyde	CC(C)(C)C1=CC=C(C=O)C=C1	13.3754363	185.568277	1	2.48E+03
13402-08-9	Acetylpheneturide	CCC(C(=O)NC(=O)NC(C)=O)C1=CC=CC=C1	64.6863655	185.56764	1	1.20E+04
3613-30-7	7-Methoxy-3,7-dimethyloctanal	COC(C)(C)CCCC(C)CC=O	32.371877	184.974049	1	5.99E+03
1078-19-9	6-Methoxy-?1-?tetralone	COC1=CC2=C(C=C1)C(=O)CCC2	40.6448301	182.567742	1	7.42E+03
5437-98-9	N-(4-Methoxyphenyl)-3-oxobutanamide	COC1=CC=C(NC(=O)CC(C)=O)C=C1	20.8415865	180.849352	1	3.77E+03
7549-33-9	Anisyl propionate	CCC(=O)OCC1=CC=C(OC)C=C1	22.8614139	180.584387	1	4.13E+03
59-63-2	Isocarboxazid	CC1=CC(=NO1)C(=O)NNCC1=CC=CC=C1	53.4387329	180.213884	1	9.63E+03
105-85-1	3,7-Dimethyloct-6-en-1-yl formate	CC(CCOC=O)CCC=C(C)C	16.723186	180.127486	1	3.01E+03
103-36-6	Ethyl cinnamate	CCOC(=O)C=CC1=CC=CC=C1	21.8081192	176.663813	1	3.85E+03
705-60-2	(2-Nitro-1-propenyl)benzene	CC(=CC1=CC=CC=C1)[N+]([O-])=O	22.6272079	174.053165	1	3.94E+03
103-37-7	Benzyl butyrate	CCCC(=O)OCC1=CC=CC=C1	17.1945352	172.821857	1	2.97E+03
142-09-6	Hexyl methacrylate	CCCCCCOC(=O)C(C)=C	8.263045 72	172.738374	1	1.43E+03
84803-46-3	4-(4-Chlorophenyl)piperidine-2,6-dione	ClC1=CC=C(C=C1)C1CC(=O)NC(=O)C1	61.7828948	172.270432	1	1.06E+04
72420-38-3	Acifran	CC1(OC(=CC1=O)C(O)=O)C1=CC=CC=C1	60.3905857	170.914397	1	1.03E+04
94-25-7	Butyl 4-aminobenzoate	CCCCOC(=O)C1=CC=C(N)C=C1	20.665683	169.560286	1	3.50E+03
306-20-7	Fenaclon	ClCCC(=O)NCCC1=CC=CC=C1	24.0809471	166.553313	1	4.01E+03
610-96-8	Methyl 2-chlorobenzoate	COC(=O)C1=C(Cl)C=CC=C1	11.6395031	163.896214	1	1.91E+03

2905-65-9	Methyl 3-chlorobenzoate	COC(=O)C1=CC(Cl)=CC=C1	10.4542221	161.924823	1	1.69E+03
7756-96-9	Butyl anthranilate	CCCCOC(=O)C1=CC=CC=C1N	15.0523676	161.091201	1	2.42E+03
77-21-4	Glutethimide	CCC1(CCC(=O)NC1=O)C1=CC=CC=C1	63.6989044	160.49966	1	1.02E+04
7335-26-4	Ethyl 2-methoxybenzoate	CCOC(=O)C1=C(OC)C=CC=C1	15.6010263	160.091994	1	2.50E+03
607-90-9	Propyl salicylate	CCCOC(=O)C1=CC=CC=C1O	16.3221064	159.674537	1	2.61E+03
125-84-8	Aminoglutethimide	CCC1(CCC(=O)NC1=O)C1=CC=C(N)C=C1	58.591081	158.44009	1	9.28E+03
48145-04-6	2-Phenoxyethyl acrylate	C=CC(=O)OCCOC1=CC=CC=C1	27.273378	156.50653	1	4.27E+03
67028-40-4	(4-Methylphenoxy) acetic acid ethyl ester	CCOC(=O)COC1=CC=C(C)C=C1	22.9942652	156.131483	1	3.59E+03
2438-72-4	Bufexamac	CCCCOC1=CC=C(CC(=O)NO)C=C1	63.0013507	150.917702	1	9.51E+03
94-02-0	Ethyl benzoylacetate	CCOC(=O)CC(=O)C1=CC=CC=C1	23.2255783	150.364736	1	3.49E+03
28315-93-7	5-Hydroxy-1-tetralone	OC1=CC=CC2=C1CCCC2=O	14.0456436	150.103938	1	2.11E+03
118-91-2	2-Chlorobenzoic acid	OC(=O)C1=CC=CC=C1Cl	14.7886005	145.770084	1	2.16E+03
78218-09-4	Dazoxiben	OC(=O)C1=CC=C(OCCN2C=CN=C2)C=C1	68.8723141	144.518479	1	9.95E+03
105-86-2	(2E)-3,7-Dimethylocta-2,6-dien-1-yl formate	CC(C)=CCC\C(C)=C\COC=O	15.3086326	143.350004	1	2.19E+03
502-47-6	Citronellic acid	CC(CCC=C(C)C)CC(O)=O	14.8366304	142.917455	1	2.12E+03
13912-80-6	Nicoboxil	CCCCOCCOC(=O)C1=CC=CN=C1	30.2356987	140.788576	1	4.26E+03
103-54-8	3-Phenylprop-2-en-1-yl acetate	CC(=O)OCC=CC1=CC=CC=C1	23.8766849	140.560148	1	3.36E+03
2315-68-6	Propyl benzoate	CCCOC(=O)C1=CC=CC=C1	10.4237502	139.087438	1	1.45E+03
104-20-1	4-(4-Methoxyphenyl)-2-butanone	COC1=CC=C(CCC(C)=O)C=C1	20.8429077	139.063817	1	2.90E+03
2021-28-5	Ethyl hydrocinnamate	CCOC(=O)CCC1=CC=CC=C1	19.869666	138.891007	1	2.76E+03
17630-75-0	5-Chlorooxindole	ClC1=CC2=C(NC(=O)C2)C=C1	11.548408	136.171615	1	1.57E+03
2050-43-3	N-(2,4-Dimethylphenyl)acetamide	CC(=O)NC1=CC=C(C)C=C1C	16.003687	136.059388	1	2.18E+03
501-68-8	Beclamide	ClCCC(=O)NCC1=CC=CC=C1	26.0219679	134.039675	1	3.49E+03
21722-83-8	2-Cyclohexylethyl acetate	CC(=O)OCCC1CCCCC1	12.2135217	133.991954	1	1.64E+03
5579-78-2	epsilon-Decalactone	CCCCC1CCCCC(=O)O1	12.0235479	132.015169	1	1.59E+03
64920-29-2	Ethyl 2-oxo-4-phenylbutyrate	CCOC(=O)C(=O)CCC1=CC=CC=C1	22.2503372	127.355848	1	2.83E+03
93-92-5	(+/-)-alpha-Methylbenzyl acetate	CC(OC(C)=O)C1=CC=CC=C1	17.9874712	127.283882	1	2.29E+03
7413-36-7	Nifenalol	CC(C)NCC(O)C1=CC=C(C=C1)[N+]([O-])=O	52.281244	126.713072	1	6.62E+03
2438-05-3	4-Propylbenzoic acid	CCCC1=CC=C(C=C1)C(O)=O	14.0960772	124.138323	1	1.75E+03
67914-60-7	4-(4-Acetylpiperazin-4-yl)phenol	CC(=O)N1CCN(CC1)C1=CC=C(O)C=C1	47.654462	124.126423	1	5.92E+03
86-35-1	Ethotoin	CCN1C(=O)NC(C1=O)C1=CC=CC=C1	62.9932444	121.208486	1	7.64E+03
459-80-3	Geranic acid	CC(C)=CCCC(C)=CC(O)=O	15.9361739	121.077571	1	1.93E+03

121-39-1	Ethyl 3-phenylglycidate	CCOC(=O)C1OC1C1=CC=CC=C1	20.2954018	120.831811	1	2.45E+03
4861-85-2	Isopropyl phenylacetate	CC(C)OC(=O)CC1=CC=CC=C1	19.5641559	119.301193	1	2.33E+03
701-64-4	Phenyl dihydrogen phosphate	OP(O)(=O)OC1=CC=CC=C1	13.3417677	119.188343	1	1.59E+03
90-49-3	Ethylphenylacetylurea	CCC(C(=O)NC(N)=O)C1=CC=CC=C1	27.9283751	118.167569	1	3.30E+03
93-68-5	N-(2-Methylphenyl)-3-oxobutanamide	CC(=O)CC(=O)NC1=C(C)C=CC=C1	25.1017458	116.943294	1	2.94E+03
80-39-7	N-Ethyl-4-methylbenzenesulfonamide	CCNS(=O)(=O)C1=CC=C(C)C=C1	21.0011897	116.923771	1	2.46E+03
480-63-7	2,4,6-Trimethylbenzoic acid	CC1=CC(C)=C(C(O)=O)C(C)=C1	13.9587216	116.276684	1	1.62E+03
1754-62-7	Methyl (E)-cinnamate	COC(=O)\C=C\C1=CC=CC=C1	16.9930887	115.027993	1	1.95E+03
103-26-4	Methyl cinnamate	COC(=O)C=CC1=CC=CC=C1	16.9930887	115.027993	1	1.95E+03
939-48-0	Propan-2-yl benzoate	CC(C)OC(=O)C1=CC=CC=C1	12.714982	114.986686	1	1.46E+03
609-66-5	2-Chlorobenzamide	NC(=O)C1=CC=CC=C1Cl	14.3388114	113.379343	1	1.63E+03
392-12-1	Indole-3-pyruvic acid	OC(=O)C(=O)CC1=CNC2=C1C=CC=C2	57.4846533	113.176955	1	6.51E+03
1009-61-6	1,1-(1,4-Phenylene)bis-ethanone	CC(=O)C1=CC=C(C=C1)C(C)=O	19.383161	112.926399	1	2.19E+03
13255-50-0	N-isopropyl-4-formylbenzamide	CC(C)NC(=O)C1=CC=C(C=O)C=C1	20.4457493	112.394644	1	2.30E+03
587-65-5	2-Chloro-N-phenylacetamide	ClCC(=O)NC1=CC=CC=C1	9.8624 8259	111.641188	1	1.10E+03
23249-97-0	Procodazole	OC(=O)CCC1=NC2=C(N1)C=CC=C2	34.0536009	111.575353	1	3.80E+03
122-72-5	3-Phenylpropyl acetate	CC(=O)OCCCC1=CC=CC=C1	21.0290393	109.434481	1	2.30E+03
27593-23-3	6-Pentyl-2H-pyran-2-one	CCCCCC1=CC=CC(=O)O1	14.2411091	108.701809	1	1.55E+03
89-25-8	1-Phenyl-3-methyl-5-pyrazolone	CC1=NN(C(=O)C1)C1=CC=CC=C1	18.4970965	105.087664	1	1.94E+03
7493-63-2	Allyl anthranilate	NC1=C(C=CC=C1)C(=O)OCC=C	18.3857277	104.098814	1	1.91E+03
98-69-1	4-Ethylbenzenesulfonic acid	CCC1=CC=C(C=C1)S(O)(=O)=O	19.1388667	102.394457	1	1.96E+03
673-31-4	Phenprobamate	NC(=O)OCCCC1=CC=CC=C1	24.9261568	101.470875	1	2.53E+03
39512-49-7	4-(4-Chlorophenyl)-4-piperidinol	OC1(CCNCC1)C1=CC=C(Cl)C=C1	39.2379299	98.8761339	1	3.88E+03
94-08-6	Ethyl 4-methylbenzoate	CCOC(=O)C1=CC=C(C)C=C1	9.9353 9485	98.7571336	1	9.81E+02
830-89-7	Albutoin	CC(C)CC1NC(=S)N(CC=C)C1=O	32.9461338	98.2607808	1	3.24E+03
5153-67-3	(E)-beta-Nitrostyrene	[O-][N+](=O)\C=C\C1=CC=CC=C1	17.3072402	96.8566429	1	1.68E+03
102-96-5	beta-Nitrostyrene	[O-][N+](=O)C=CC1=CC=CC=C1	16.673597	96.8566429	1	1.61E+03
15351-09-4	Metamfepramone	CC(N(C)C)C(=O)C1=CC=CC=C1	16.5585395	92.2510742	1	1.53E+03
7424-00-2	Fenclonine	NC(CC1=CC=C(Cl)C=C1)C(O)=O	28.5523804	90.2580009	1	2.58E+03
4350-09-8	L-5-Hydroxytryptophan	N[C@@H](CC1=CNC2=C1C=C(O)C=C2)C(O)=O	73.3573756	88.9849458	1	6.53E+03
87-24-1	Ethyl 2-methylbenzoate	CCOC(=O)C1=C(C)C=CC=C1	10.8690743	87.275159	1	9.49E+02
2244-16-8	d-Carvone	CC(=C)[C@H]1CC=C(C)C(=O)C1	8.6166 3748	84.9810636	1	7.32E+02

6485-40-1	R-(-)-Carvone	CC(=C)[C@@H]1CC=C(C)C(=O)C1	8.6166 3748	84.981 0636	1	7.32E +02
99-49-0	dl-Carvone	CC(=C)C1CC=C(C)C(=O)C1	8.6166 3748	84.981 0636	1	7.32E +02
111-80-8	Methyl 2-nonynoate	CCCCCCC#CC(=O)OC	8.0275 3548	83.568 8862	1	6.71E +02
536-69-6	Fusaric acid	CCCCC1=CN=C(C=C1)C(O)=O	17.223 8243	82.856 8035	1	1.43E +03
104-21-2	4-Methoxybenzyl acetate	COC1=CC=C(COC(C)=O)C=C1	27.015 4147	82.339 0104	1	2.22E +03
54-12-6	dl-Tryptophan	NC(CC1=CNC2=CC=CC=C12)C(O)=O	67.833 7502	79.062 3143	1	5.36E +03
73-22-3	l-Tryptophan	N[C@@H](CC1=CNC2=CC=CC=C12)C(O)=O	67.833 7502	79.062 3143	1	5.36E +03
153-94-6	d-Tryptophan	N[C@H](CC1=CNC2=CC=CC=C12)C(O)=O	67.833 7502	79.062 3143	1	5.36E +03
5471-51-2	4-(4-Hydroxyphenyl)butan-2-one	CC(=O)CCC1=CC=C(O)C=C1	21.451 8619	76.950 8728	1	1.65E +03
140-39-6	4-Tolyl acetate	CC(=O)OC1=CC=C(C)C=C1	10.976 9384	76.628 8434	1	8.41E +02
2941-55-1	Ethiolate	CCSC(=O)N(CC)CC	10.669 2904	75.940 3422	1	8.10E +02
7473-98-5	Propylene glycol diacetate	CC(C)(O)C(=O)C1=CC=CC=C1	13.990 2806	73.606 2712	1	1.03E +03
86-34-0	Phensuximide	CN1C(=O)CC(C1=O)C1=CC=CC=C1	66.556 2531	70.536 3091	1	4.69E +03
101-97-3	Ethyl phenylacetate	CCOC(=O)CC1=CC=CC=C1	16.247 0523	70.517 0132	1	1.15E +03
6837-24-7	1-Cyclohexylpyrrolidin-2-one	O=C1CCCN1C1CCCCC1	27.978 7576	69.568 317	1	1.95E +03
60397-77-5	N-(2,4-Dimethylphenyl)formamide	CC1=CC=C(NC=O)C(C)=C1	8.0105 2072	68.556 8864	1	5.49E +02
99-75-2	Methyl 4-methylbenzoate	COC(=O)C1=CC=C(C)C=C1	8.2316 7306	67.771 5091	1	5.58E +02
30709-69-4	Tizoprolic acid	CCCC1=NC=C(S1)C(O)=O	9.2852 2602	67.663 9554	1	6.28E +02
103-89-9	N-Acetyl-p-toluidine	CC(=O)NC1=CC=C(C)C=C1	12.541 4596	67.240 7141	1	8.43E +02
4822-44-0	Thioglycolic acid anilide	SCC(=O)NC1=CC=CC=C1	9.5146 4123	65.423 2075	1	6.22E +02
774-40-3	Ethyl mandelate	CCOC(=O)C(O)C1=CC=CC=C1	22.612 5742	63.915 7068	1	1.45E +03
537-55-3	N-Acetyl-L-tyrosine	CC(=O)N[C@@H](CC1=CC=C(O)C=C1)C(O)=O	46.721 9608	61.216 6297	1	2.86E +03
1878-49-5	(2-Methylphenoxy)acetic acid	CC1=C(OCC(O)=O)C=CC=C1	25.369 2383	59.991 5398	1	1.52E +03
103-45-7	2-Phenylethyl acetate	CC(=O)OCCC1=CC=CC=C1	19.293 2673	57.650 4214	1	1.11E +03
122-46-3	m-Cresyl acetate	CC(=O)OC1=CC=CC(C)=C1	10.307 8302	52.743 7118	1	5.44E +02
2901-75-9	Afalanine	CC(=O)NC(CC1=CC=CC=C1)C(O)=O	21.197 8643	51.998 9145	1	1.10E +03
1701-77-5	Methoxyphenylacetic acid	COC(C(O)=O)C1=CC=CC=C1	16.748 3207	50.844 3174	1	8.52E +02
6961-46-2	Idrocilamide	OCCNC(=O)C=CC1=CC=CC=C1	25.660 3137	47.663 2044	1	1.22E +03
120-66-1	N-Acetyl-o-toluidine	CC(=O)NC1=C(C)C=CC=C1	10.508 5029	47.097 8152	1	4.95E +02
15121-84-3	2-(2-Nitrophenyl)ethanol	OCCC1=C(C=CC=C1)[N+]([O-])=O	23.518 6045	43.449 9093	1	1.02E +03
537-92-8	N-Acetyl-m-toluidine	CC(=O)NC1=CC=CC(C)=C1	11.719 8664	43.323 3535	1	5.08E +02
100-27-6	2-(4-Nitrophenyl)ethanol	OCCC1=CC=C(C=C1)[N+]([O-])=O	20.238 9172	41.299 8418	1	8.36E +02

156-06-9	Phenylpyruvic acid	OC(=O)C(=O)CC1=CC=CC=C1	25.78592	37.8249695	1	9.75E+02
15302-18-8	Formetorex	CC(CC1=CC=CC=C1)NC=O	17.4643932	37.3287919	1	6.52E+02
5251-93-4	Benzadox	OC(=O)CONC(=O)C1=CC=CC=C1	23.7738089	36.2084335	1	8.61E+02
63721-05-1	Methyl 3,3-dimethylpent-4-enoate	COC(=O)CC(C)(C)C=C	8.32633379	32.3025067	1	2.69E+02

Reference

(1) Finnveden, G.; Hauschild, M. Z.; Ekvall, T.; Guinée, J.; Heijungs, R.; Hellweg, S.; Koehler, A.; Pennington, D.; Suh, S. Recent Developments in Life Cycle Assessment. *Journal of Environmental Management* **2009**, *91* (1), 1–21. https://doi.org/10.1016/j.jenvman.2009.06.018.

(2) Standardization, I. O. for. *Environmental Management: Life Cycle Assessment; Principles and Framework*; ISO, 2006.

(3) Alshamrani, O. S.; Galal, K.; Alkass, S. Integrated LCA–LEED Sustainability Assessment Model for Structure and Envelope Systems of School Buildings. *Building and Environment* **2014**, *80*, 61–70.

(4) Hunt, R. G.; Franklin, W. E.; Hunt, R. G. LCA—How It Came About. *The international journal of life cycle assessment* **1996**, *1* (1), 4–7.

(5) Guinée, J. B.; Heijungs, R.; Huppes, G.; Zamagni, A.; Masoni, P.; Buonamici, R.; Ekvall, T.; Rydberg, T. Life Cycle Assessment: Past, Present, and Future. *Environ. Sci. Technol.* **2011**, *45* (1), 90–96. https://doi.org/10.1021/es101316v.

(6) Heijungs, R.; Suh, S. *The Computational Structure of Life Cycle Assessment*; Springer Science & Business Media, 2013; Vol. 11.

(7) Guinee, J. B. Handbook on Life Cycle Assessment Operational Guide to the ISO Standards. *Int J LCA* **2002**, *7* (5), 311–313. https://doi.org/10.1007/BF02978897.

(8) Rosenbaum, R. K.; Huijbregts, M. A. J.; Henderson, A. D.; Margni, M.; McKone, T. E.; Meent, D. van de; Hauschild, M. Z.; Shaked, S.; Li, D. S.; Gold, L. S.; et al. USEtox Human Exposure and Toxicity Factors for Comparative Assessment of Toxic Emissions in Life Cycle Analysis: Sensitivity to Key Chemical Properties. *Int J Life Cycle Assess* **2011**, *16* (8), 710. https://doi.org/10.1007/s11367-011-0316-4.

(9) MCKONE, E. G. H. W. S. P. T. E. Environmental Policy Analysis: Evaluating Toxic Impact Assessment Methods: What Works Best. *Environmental science & technology* **1998**, *32* (5), 138A-144A.

(10) Huijbregts, M. A. J.; Struijs, J.; Goedkoop, M.; Heijungs, R.; Jan Hendriks, A.; van de Meent, D. Human Population Intake Fractions and Environmental Fate Factors of Toxic Pollutants in Life Cycle Impact Assessment. *Chemosphere* **2005**, *61* (10), 1495–1504. https://doi.org/10.1016/j.chemosphere.2005.04.046.

(11) Hischier, R.; Hellweg, S.; Capello, C.; Primas, A. Establishing Life Cycle Inventories of Chemicals Based on Differing Data Availability. *Int J Life Cycle Assessment* **2005**, *10* (1), 59–67. https://doi.org/10.1065/lca2004.10.181.7.

(12) 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.; et al. USEtox—the UNEP-SETAC Toxicity Model: Recommended Characterisation Factors for Human Toxicity and Freshwater Ecotoxicity in Life Cycle Impact Assessment. *Int J Life Cycle Assess* **2008**, *13* (7), 532–546. https://doi.org/10.1007/s11367-008-0038-4.

(13) 2.10.2 Direct Global Warming Potentials - AR4 WGI Chapter 2: Changes in Atmospheric Constituents and in Radiative Forcing https://www.ipcc.ch/publications_and_data/ar4/wg1/en/ch2s2-10-2.html (accessed Apr 26, 2017).

(14) Empowering Innovation & Scientific Discoveries | CAS https://www.cas.org/ (accessed Dec 30, 2018).
(15) Efremenkova, V. M.; Krukovskaya, N. V. Chemical Abstracts Service Centennial: Facts and Figures. *Sci. Tech.Inf. Proc.* **2007**, *34* (6), 328–334. https://doi.org/10.3103/S0147688207060093.
(16) Marengo, E.; Bobba, M.; Robotti, E.; Liparota, M. C. Modeling of the Polluting Emissions from a Cement Production Plant by Partial Least-Squares, Principal Component Regression, and Artificial Neural Networks. *Environ. Sci. Technol.* **2005**, *40* (1), 272–280. https://doi.org/10.1021/es0517466.
(17) Park, J. H.; Seo, K.-K. Approximate Life Cycle Assessment of Product Concepts Using Multiple Regression Analysis and Artificial Neural Networks. *KSME International Journal* **2003**, *17* (12), 1969–1976. https://doi.org/10.1007/BF02982436.
(18) Pascual-González, J.; Pozo, C.; Guillén-Gosálbez, G.; Jiménez-Esteller, L. Combined Use of MILP and Multi-Linear Regression to Simplify LCA Studies. *Comput. Chem. Eng.* **2015**, *82*, 34–43. https://doi.org/10.1016/j.compchemeng.2015.06.002.
(19) Allison, T. C. Application of an Artificial Neural Network to the Prediction of OH Radical Reaction Rate Constants for Evaluating Global Warming Potential. *J. Phys. Chem. B* **2016**, *120* (8), 1854–1863. https://doi.org/10.1021/acs.jpcb.5b09558.
(20) Andersson, K.; Ohlsson, T.; Olsson, P. Screening Life Cycle Assessment (LCA) of Tomato Ketchup: A Case Study. *Journal of Cleaner Production* **1998**, *6* (3–4), 277–288. https://doi.org/10.1016/S0959-6526(98)00027-4.
(21) Carriger, J. F.; Martin, T. M.; Barron, M. G. A Bayesian Network Model for Predicting Aquatic Toxicity Mode of Action Using Two Dimensional Theoretical Molecular Descriptors. *Aquatic Toxicology* **2016**, *180*, 11–24. https://doi.org/10.1016/j.aquatox.2016.09.006.
(22) Sousa, I.; Wallace, D.; Eisenhard, J. L. Approximate Life-Cycle Assessment of Product Concepts Using Learning Systems. *Journal of Industrial Ecology* **2000**, *4* (4), 61–81. https://doi.org/10.1162/10881980052541954.
(23) Wernet, G.; Papadokonstantakis, S.; Hellweg, S.; Hungerbühler, K. Bridging Data Gaps in Environmental Assessments: Modeling Impacts of Fine and Basic Chemical Production. *Green Chem.* **2009**, *11* (11), 1826–1831. https://doi.org/10.1039/B905558D.
(24) Glahn, H. R.; Lowry, D. A. The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *J. Appl. Meteor.* **1972**, *11* (8), 1203–1211. https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.
(25) Maier, H. R.; Dandy, G. C. Neural Networks for the Prediction and Forecasting of Water Resources Variables: A Review of Modelling Issues and Applications. *Environmental Modelling & Software* **2000**, *15* (1), 101–124. https://doi.org/10.1016/S1364-8152(99)00007-9.
(26) Hsu, K.; Gupta, H. V.; Sorooshian, S. Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resour. Res.* **1995**, *31* (10), 2517–2530. https://doi.org/10.1029/95WR01955.

(27) Walter, A.; Denhard, M.; Schönwiese, C.-D. Simulation of Global Temperature Variations and Signal Detection Studies Using Neural Networks. *MyScienceWork* **2005**.
(28) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokonstantakis, S.; Hungerbühler, K. Molecular-Structure-Based Models of Chemical Inventories Using Neural Networks. *Environ. Sci. Technol.* **2008**, *42* (17), 6717–6722. https://doi.org/10.1021/es7022362.
(29) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Current Pharmaceutical Design* **2007**, *13* (34), 3494–3504. https://doi.org/10.2174/138161207782794257.
(30) Joanna Jaworska, N. N.-J. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *Altern. Lab Anim.* **2005**, *33* (5), 445–459.
(31) Alternatives Analysis Guide; https://www.dtsc.ca.gov/SCP/AlternativesAnalysisGuidance.cfm.
(32) Tabone, M. D.; Cregg, J. J.; Beckman, E. J.; Landis, A. E. Sustainability Metrics: Life Cycle Assessment and Green Design in Polymers. *Environ. Sci. Technol.* **2010**, *44* (21), 8264–8269. https://doi.org/10.1021/es101640n.
(33) Eckelman, M. J. Life Cycle Inherent Toxicity: A Novel LCA-Based Algorithm for Evaluating Chemical Synthesis Pathways. *Green Chem.* **2016**. https://doi.org/10.1039/C5GC02768C.
(34) Anastas, P. T.; Lankey, R. L. Life Cycle Assessment and Green Chemistry: The Yin and Yang of Industrial Ecology. *Green Chem.* **2000**, *6*, 289–295. https://doi.org/10.1039/B005650M.
(35) Candidate Chemical List https://www.dtsc.ca.gov/SCP/CandidateChemicals.cfm (accessed Jan 22, 2016).
(36) Hischier, R.; Hellweg, S.; Capello, C.; Primas, A. Establishing Life Cycle Inventories of Chemicals Based on Differing Data Availability. *Int J Life Cycle Assess* **2004**, *10* (1), 59–67. https://doi.org/10.1065/lca2004.10.181.7.
(37) Verghese, K. L.; Horne, R.; Carre, A. PIQET: The Design and Development of an Online 'Streamlined' LCA Tool for Sustainable Packaging Design Decision Support. *Int J Life Cycle Assess* **2010**, *15* (6), 608–620. https://doi.org/10.1007/s11367-010-0193-2.
(38) Jiménez-González, C.; Constable, D. J. C.; Ponder, C. S. Evaluating the "Greenness" of Chemical Processes and Products in the Pharmaceutical Industry—a Green Metrics Primer. *Chem. Soc. Rev.* **2012**, *41* (4), 1485–1498. https://doi.org/10.1039/C1CS15215G.
(39) Roches, A.; Nemecek, T.; Gaillard, G.; Plassmann, K.; Sim, S.; King, H.; Canals, L. M. i. MEXALCA: A Modular Method for the Extrapolation of Crop LCA. *Int J Life Cycle Assess* **2010**, *15* (8), 842–854. https://doi.org/10.1007/s11367-010-0209-y.
(40) Bala, A.; Raugei, M.; Benveniste, G.; Gazulla, C.; Fullana-i-Palmer, P. Simplified Tools for Global Warming Potential Evaluation: When 'Good Enough' Is Best. *Int J Life Cycle Assess* **2010**, *15* (5), 489–498. https://doi.org/10.1007/s11367-010-0153-x.

(41) Casamayor, J. L.; Su, D. Integration of Detailed/Screening LCA Software-Based Tools into Design Processes. In *Design for Innovative Value Towards a Sustainable Society*; Matsumoto, D. M., Umeda, P. Y., Masui, D. K., Fukushige, D. S., Eds.; Springer Netherlands, 2012; pp 609–614. https://doi.org/10.1007/978-94-007-3010-6_117.
(42) Canals, L. M. i; Azapagic, A.; Doka, G.; Jefferies, D.; King, H.; Mutel, C.; Nemecek, T.; Roches, A.; Sim, S.; Stichnothe, H.; et al. Approaches for Addressing Life Cycle Assessment Data Gaps for Bio-Based Products. *J. Ind. Ecol.* **2011**, *15* (5), 707–725. https://doi.org/10.1111/j.1530-9290.2011.00369.x.
(43) Steinmann, Z. J. N.; Venkatesh, A.; Hauck, M.; Schipper, A. M.; Karuppiah, R.; Laurenzi, I. J.; Huijbregts, M. A. J. How To Address Data Gaps in Life Cycle Inventories: A Case Study on Estimating CO2 Emissions from Coal-Fired Electricity Plants on a Global Scale. *Environ. Sci. Technol.* **2014**, *48* (9), 5282–5289. https://doi.org/10.1021/es500757p.
(44) Hanes, R.; Bakshi, B. R.; Goel, P. K. The Use of Regression in Streamlined Life Cycle Assessment. In *International Symposium on Sustainable Systems and Technologies, at Cincinnati, OH*; 2013.
(45) Niero, M.; Felice, F. D.; Ren, J.; Manzardo, A.; Scipioni, A. How Can a Life Cycle Inventory Parametric Model Streamline Life Cycle Assessment in the Wooden Pallet Sector? *Int J Life Cycle Assess* **2014**, *19* (4), 901–918. https://doi.org/10.1007/s11367-014-0705-6.
(46) Igor I Baskin, V. A. P. Neural Networks in Building QSAR Models. *Methods in molecular biology (Clifton, N.J.)* **2008**, *458*, 133–154. https://doi.org/10.1007/978-1-60327-101-1_8.
(47) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.* **2010**, *110* (10), 5714–5789. https://doi.org/10.1021/cr900238d.
(48) Xiao, R.; Ye, T.; Wei, Z.; Luo, S.; Yang, Z.; Spinney, R. Quantitative Structure–Activity Relationship (QSAR) for the Oxidation of Trace Organic Contaminants by Sulfate Radical. *Environ. Sci. Technol.* **2015**, *49* (22), 13394–13402. https://doi.org/10.1021/acs.est.5b03078.
(49) Vink, E. T. H.; Rábago, K. R.; Glassner, D. A.; Gruber, P. R. Applications of Life Cycle Assessment to NatureWorks™ Polylactide (PLA) Production. *Polym. Degrad. Stab.* **2003**, *80* (3), 403–419. https://doi.org/10.1016/S0141-3910(02)00372-5.
(50) Assen, N. von der; Bardow, A. Life Cycle Assessment of Polyols for Polyurethane Production Using CO 2 as Feedstock: Insights from an Industrial Case Study. *Green Chem.* **2014**, *16* (6), 3272–3280. https://doi.org/10.1039/C4GC00513A.
(51) Wernet, G.; Conradt, S.; Isenring, H. P.; Jiménez-González, C.; Hungerbühler, K. Life Cycle Assessment of Fine Chemical Production: A Case Study of Pharmaceutical Synthesis. *Int J Life Cycle Assess* **2010**, *15* (3), 294–303. https://doi.org/10.1007/s11367-010-0151-z.
(52) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokonstantakis, S.; Hungerbühler, K. Molecular-Structure-Based Models of Chemical Inventories Using Neural

Networks. *Environ. Sci. Technol.* **2008**, *42* (17), 6717–6722. https://doi.org/10.1021/es7022362.
(53) Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR Comb. Sci.* **2007**, *26* (5), 694–701. https://doi.org/10.1002/qsar.200610151.
(54) Validation of (Q)SAR Models - OECD; http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed Apr 26, 2016).
(55) Kalogirou, S. A. Artificial Neural Networks in Renewable Energy Systems Applications: A Review. *Renew. Sustainable Energy Rev.* **2001**, *5* (4), 373–401. https://doi.org/10.1016/S1364-0321(01)00006-5.
(56) Gardner, M. W.; Dorling, S. R. Artificial Neural Networks (the Multilayer Perceptron)—a Review of Applications in the Atmospheric Sciences. *Atmos. Environ.* **1998**, *32* (14–15), 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0.
(57) Hornik, K.; Stinchcombe, M.; White, H. Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks* **1989**, *2* (5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.
(58) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **2015**, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.
(59) Hanrahan, G. *Artificial Neural Networks in Biological and Environmental Analysis*; CRC Press, 2011.
(60) Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A. r; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* **2012**, *29* (6), 82–97. https://doi.org/10.1109/MSP.2012.2205597.
(61) Hanrahan, G. *Artificial Neural Networks in Biological and Environmental Analysis*; CRC Press, 2011.
(62) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:1603.04467* **2016**.
(63) Weidema, B.P.; Bauer, Ch.; Hischier, R.; Mutel, Ch.; Nemecek, T.; Reinhard, J.; Vadenbo, C.O.; Wernet, G. The Ecoinvent Database: Overview and Methodology, Data Quality Guideline for the Ecoinvent Database Version 3. *www.ecoinvent.org*.
(64) Huijbregts, M. A. J.; Hellweg, S.; Frischknecht, R.; Hendriks, H. W. M.; Hungerbühler, K.; Hendriks, A. J. Cumulative Energy Demand As Predictor for the Environmental Burden of Commodity Production. *Environ. Sci. Technol.* **2010**, *44* (6), 2189–2196. https://doi.org/10.1021/es902870s.
(65) IPCC. *Climate Change 2007: The Physical Science Basis*; Climate Change 2007; Cambridge University Press: Cambridge, UK.
(66) Bare, J. TRACI 2.0: The Tool for the Reduction and Assessment of Chemical and Other Environmental Impacts 2.0. *Clean Techn Environ Policy* **2011**, *13* (5), 687–696. https://doi.org/10.1007/s10098-010-0338-9.
(67) Goedkoop, M.; Spriensma, R. *The Eco-Indicator99: A Damage Oriented Method for Life Cycle Impact Assessment: Methodology Report*; 2001; pp 1–144.
(68) Jolliet, O.; Margni, M.; Charles, R.; Humbert, S.; Payet, J.; Rebitzer, G.; Rosenbaum, R. IMPACT 2002+: A New Life Cycle Impact Assessment

Methodology. *Int J LCA* **2003**, *8* (6), 324–330. https://doi.org/10.1007/BF02978505.
(69) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience, 2000.
(70) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graph. Model.* **2000**, *18* (4–5), 464–477. https://doi.org/10.1016/S1093-3263(00)00068-1.
(71) Kode - Chemoinformatics https://chm.kode-solutions.net/products_dragon.php (accessed Apr 26, 2017).
(72) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.
(73) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010. https://doi.org/10.1021/jm4004285.
(74) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
(75) Lauria, A.; Ippolito, M.; Almerico, A. M. Combined Use of PCA and QSAR/QSPR to Predict the Drugs Mechanism of Action. An Application to the NCI ACAM Database. *QSAR Comb. Sci.* **2009**, *28* (4), 387–395. https://doi.org/10.1002/qsar.200810062.
(76) Khan, J.; Wei, J. S.; Ringnér, M.; Saal, L. H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C. R.; Peterson, C.; et al. Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks. *Nat Med* **2001**, *7* (6), 673–679. https://doi.org/10.1038/89044.
(77) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313* (5786), 504–507. https://doi.org/10.1126/science.1127647.
(78) Saeys, Y.; Inza, I.; Larrañaga, P. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics* **2007**, *23* (19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344.
(79) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How Not to Develop a Quantitative Structure–Activity or Structure–Property Relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research* **2009**, *20* (3–4), 241–266. https://doi.org/10.1080/10629360902949567.
(80) Mosteller, F.; Tukey, J. W. Data Analysis, Including Statistics. **1968**.
(81) Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *Altern Lab Anim* **2005**, *33* (2), 155–173.
(82) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 1912–1928. https://doi.org/10.1021/ci049782w.

(83) Hartley, M. J. Rationale and Methods for Conserving Biodiversity in Plantation Forests. *Forest Ecology and Management* **2002**, *155* (1), 81–95. https://doi.org/10.1016/S0378-1127(01)00549-7.
(84) Vörösmarty, C. J.; McIntyre, P. B.; Gessner, M. O.; Dudgeon, D.; Prusevich, A.; Green, P.; Glidden, S.; Bunn, S. E.; Sullivan, C. A.; Liermann, C. R.; et al. Global Threats to Human Water Security and River Biodiversity. *Nature* **2010**, *467* (7315), 555–561. https://doi.org/10.1038/nature09440.
(85) Malaj, E.; Ohe, P. C. von der; Grote, M.; Kühne, R.; Mondy, C. P.; Usseglio-Polatera, P.; Brack, W.; Schäfer, R. B. Organic Chemicals Jeopardize the Health of Freshwater Ecosystems on the Continental Scale. *PNAS* **2014**, *111* (26), 9549–9554. https://doi.org/10.1073/pnas.1321082111.
(86) The IUCN Red List of Threatened Species http://www.iucnredlist.org/ (accessed Jun 10, 2018).
(87) Holmstrup, M.; Bindesbøl, A.-M.; Oostingh, G. J.; Duschl, A.; Scheil, V.; Köhler, H.-R.; Loureiro, S.; Soares, A. M. V. M.; Ferreira, A. L. G.; Kienle, C.; et al. Interactions between Effects of Environmental Chemicals and Natural Stressors: A Review. *Science of The Total Environment* **2010**, *408* (18), 3746–3762. https://doi.org/10.1016/j.scitotenv.2009.10.067.
(88) Bressler, D. W.; Stribling, J. B.; Paul, M. J.; Hicks, M. B. Stressor Tolerance Values for Benthic Macroinvertebrates in Mississippi. *Hydrobiologia* **2006**, *573* (1), 155–172. https://doi.org/10.1007/s10750-006-0266-1.
(89) US EPA. 2016 Chemical Data Reporting.
(90) ECHA Publishes Official Statistics for the Last REACH Registration Deadline https://www.chemsafetypro.com/Topics/EU/ECHA_Publishes_2018_REACH_Registration_Statistics.html (accessed Feb 4, 2019).
(91) Wolansky, M. J.; Harrill, J. A. Neurobehavioral Toxicology of Pyrethroid Insecticides in Adult Animals: A Critical Review. *Neurotoxicology and Teratology* **2008**, *30* (2), 55–78. https://doi.org/10.1016/j.ntt.2007.10.005.
(92) forum, E.-U. E. protection agency R. assessment. *Guidelines for Ecological Risk Assessment*; US Environmental protection agency, 1998.
(93) Posthuma, L.; II, G. W. S.; Traas, T. P. *Species Sensitivity Distributions in Ecotoxicology*; CRC Press, 2001.
(94) Qi Ping; Wang Ying; Mu Jingli; Wang Juying. Aquatic Predicted No-effect-concentration Derivation for Perfluorooctane Sulfonic Acid. *Environmental Toxicology and Chemistry* **2011**, *30* (4), 836–842. https://doi.org/10.1002/etc.460.
(95) Raimondo, S.; Vivian, D. N.; Delos, C.; Barron, M. G. Protectiveness of Species Sensitivity Distribution Hazard Concentrations for Acute Toxicity Used in Endangered Species Risk Assessment. *Environmental Toxicology and Chemistry* **2008**, *27* (12), 2599–2607. https://doi.org/10.1897/08-157.1.
(96) Cunningham, V. L.; Binks, S. P.; Olson, M. J. Human Health Risk Assessment from the Presence of Human Pharmaceuticals in the Aquatic Environment. *Regulatory Toxicology and Pharmacology* **2009**, *53* (1), 39–45.
(97) Calow, P.; Forbes, V. E. *Peer Reviewed: Does Ecotoxicology Inform Ecological Risk Assessment?*; ACS Publications, 2003.
(98) 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. *Int J Life Cycle Assess* **2011**, *16* (8), 701. https://doi.org/10.1007/s11367-011-0294-6.

(99) Garner, K. L.; Suh, S.; Lenihan, H. S.; Keller, A. A. Species Sensitivity Distributions for Engineered Nanomaterials. *Environ. Sci. Technol.* **2015**, *49* (9), 5753–5759. https://doi.org/10.1021/acs.est.5b00081.

(100) Lowry, G. V.; Espinasse, B. P.; Badireddy, A. R.; Richardson, C. J.; Reinsch, B. C.; Bryant, L. D.; Bone, A. J.; Deonarine, A.; Chae, S.; Therezien, M.; et al. Long-Term Transformation and Fate of Manufactured Ag Nanoparticles in a Simulated Large Scale Freshwater Emergent Wetland. *Environ. Sci. Technol.* **2012**, *46* (13), 7027–7036. https://doi.org/10.1021/es204608d.

(101) Newman Michael C.; Ownby David R.; Mézin Laurent C. A.; Powell David C.; Christensen Tyler R. L.; Lerberg Scott B.; Anderson Britt-Anne. Applying Species-sensitivity Distributions in Ecological Risk Assessment: Assumptions of Distribution Type and Sufficient Numbers of Species. *Environmental Toxicology and Chemistry* **2009**, *19* (2), 508–515. https://doi.org/10.1002/etc.5620190233.

(102) Andersen, M. E.; Krewski, D. Toxicity Testing in the 21st Century: Bringing the Vision to Life. *Toxicol Sci* **2009**, *107* (2), 324–330. https://doi.org/10.1093/toxsci/kfn255.

(103) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010. https://doi.org/10.1021/jm4004285.

(104) Worth, A. P.; Cronin, M. T. The Use of Discriminant Analysis, Logistic Regression and Classification Tree Analysis in the Development of Classification Models for Human Health Effects. *Journal of Molecular Structure: THEOCHEM* **2003**, *622* (1–2), 97–111.

(105) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of Random Forest and Pipeline Pilot Naive Bayes in Prospective QSAR Predictions. *Journal of chemical information and modeling* **2012**, *52* (3), 792–803.

(106) Mayer, F. L.; Krause, G. F.; Ellersieck, M. R.; Lee, G.; Buckler, D. R. Predicting Chronic Lethality of Chemicals to Fishes from Acute Toxicity Test Data: Concepts and Linear Regression Analysis. *Environmental Toxicology and Chemistry* **1994**, *13* (4), 671–678. https://doi.org/10.1002/etc.5620130418.

(107) Raevsky, O. A.; Grigor'ev, V. Y.; Weber, E. E.; Dearden, J. C. Classification and Quantification of the Toxicity of Chemicals to Guppy, Fathead Minnow and Rainbow Trout: Part 1. Nonpolar Narcosis Mode of Action. *QSAR Comb. Sci.* **2008**, *27* (11–12), 1274–1281. https://doi.org/10.1002/qsar.200860014.

(108) Haupt, S. E.; Pasini, A.; Marzban, C. *Artificial Intelligence Methods in the Environmental Sciences*; Springer Science & Business Media, 2008.

(109) Liu, Y.; Racah, E.; Prabhat; Correa, J.; Khosrowshahi, A.; Lavers, D.; Kunkel, K.; Wehner, M.; Collins, W. Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets. **2016**.

(110) Devillers, J. A General QSAR Model for Predicting the Acute Toxicity of Pesticides to Lepomis Macrochirus. *SAR and QSAR in Environmental Research* **2001**, *11* (5–6), 397–417. https://doi.org/10.1080/10629360108035361.

(111) Martin, T. M.; Young, D. M. Prediction of the Acute Toxicity (96-h LC50) of Organic Compounds to the Fathead Minnow (Pimephales Promelas) Using a Group Contribution Method. *Chem. Res. Toxicol.* **2001**, *14* (10), 1378–1385. https://doi.org/10.1021/tx0155045.
(112) Kaiser, K. L. E. The Use of Neural Networks in QSARs for Acute Aquatic Toxicological Endpoints. *Journal of Molecular Structure: THEOCHEM* **2003**, *622* (1–2), 85–95. https://doi.org/10.1016/S0166-1280(02)00620-6.
(113) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42* (16), 3183–3187. https://doi.org/10.1021/jm980697n.
(114) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons, 2008.
(115) eChemPortal - Home https://www.echemportal.org/echemportal/propertysearch/page.action;jsessionid=D34DADB24143BE5071985CCDC085AA77?pageID=0 (accessed Oct 17, 2017).
(116) ECOTOX | MED | US EPA https://cfpub.epa.gov/ecotox/ecotox_home.cfm (accessed Oct 17, 2017).
(117) ESFA https://dwh.efsa.europa.eu/bi/asp/Main.aspx (accessed Oct 17, 2017).
(118) Hazardous Substances Data Bank (HSDB) https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB (accessed Oct 17, 2017).
(119) Results of Eco-Toxicity Tests Data Conducted by Ministry of the Environment in Japan, 2014.
(120) Austin, T.; Denoyelle, M.; Chaudry, A.; Stradling, S.; Eadsforth, C. European Chemicals Agency Dossier Submissions as an Experimental Data Source: Refinement of a Fish Toxicity Model for Predicting Acute LC50 Values. *Environ Toxicol Chem* **2015**, *34* (2), 369–378. https://doi.org/10.1002/etc.2817.
(121) Toropov, A. A.; Toropova, A. P.; Marzo, M.; Dorne, J. L.; Georgiadis, N.; Benfenati, E. QSAR Models for Predicting Acute Toxicity of Pesticides in Rainbow Trout Using the CORAL Software and EFSA's OpenFoodTox Database. *Environmental Toxicology and Pharmacology* **2017**, *53* (Supplement C), 158–163. https://doi.org/10.1016/j.etap.2017.05.011.
(122) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting Modes of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (Pimephales Promelas). *Environmental Toxicology and Chemistry* **1997**, *16* (5), 948–967. https://doi.org/10.1002/etc.5620160514.
(123) *Rdkit: The Official Sources for the RDKit Library*; RDKit, 2017.
(124) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *Journal of Cheminformatics* **2018**, *10*, 4. https://doi.org/10.1186/s13321-018-0258-y.
(125) Dutta, D.; Guha, R.; Wild, D.; Chen, T. Ensemble Feature Selection: Consistent Descriptor Subsets for Multiple QSAR Models. *J. Chem. Inf. Model.* **2007**, *47* (3), 989–997. https://doi.org/10.1021/ci600563w.
(126) Sugumaran, V.; Muralidharan, V.; Ramachandran, K. I. Feature Selection Using Decision Tree and Classification through Proximal Support Vector Machine for

Fault Diagnostics of Roller Bearing. *Mechanical Systems and Signal Processing* **2007**, *21* (2), 930–942. https://doi.org/10.1016/j.ymssp.2006.05.004.
(127) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12* (Oct), 2825–2830.
(128) Chollet, F. *Keras*; GitHub, 2015.
(129) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs]* **2016**.
(130) ppoints function | R Documentation https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/ppoints (accessed Feb 8, 2019).
(131) Aldenberg, T.; Rorije, E. Species Sensitivity Distribution Estimation from Uncertain (QSAR-Based) Effects Data. *Altern Lab Anim* **2013**, *41* (1), 19–31.
(132) Wheeler, J. R.; Grist, E. P. M.; Leung, K. M. Y.; Morritt, D.; Crane, M. Species Sensitivity Distributions: Data and Model Choice. *Marine Pollution Bulletin* **2002**, *45* (1), 192–202. https://doi.org/10.1016/S0025-326X(01)00327-7.
(133) Aldenberg, T.; Slob, W. Confidence Limits for Hazardous Concentrations Based on Logistically Distributed NOEC Toxicity Data. *Ecotoxicology and Environmental Safety* **1993**, *25* (1), 48–63. https://doi.org/10.1006/eesa.1993.1006.
(134) Qin, Y.; Suh, S. What Distribution Function Do Life Cycle Inventories Follow? *Int J Life Cycle Assess* **2017**, *22* (7), 1138–1145. https://doi.org/10.1007/s11367-016-1224-4.
(135) MacKinnon, D. P.; Lockwood, C. M.; Williams, J. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behavioral Research* **2004**, *39* (1), 99–128. https://doi.org/10.1207/s15327906mbr3901_4.
(136) US EPA, O. Toxicology Testing in the 21st Century (Tox21) https://www.epa.gov/chemical-research/toxicology-testing-21st-century-tox21 (accessed Mar 2, 2018).
(137) Registered substances - ECHA https://echa.europa.eu/information-on-chemicals/registered-substances (accessed Feb 13, 2019).
(138) Reap, J.; Roman, F.; Duncan, S.; Bras, B. A Survey of Unresolved Problems in Life Cycle Assessment. *The international journal of life cycle assessment* **2008**, *13* (5), 374.
(139) Randall, D.; Lee, S. *The Polyurethanes Book*; Wiley, 2002.
(140) US EPA, O. Methylene Diphenyl Diisocyanate (MDI) And Related Compounds https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/methylene-diphenyl-diisocyanate-mdi-and-related (accessed Feb 12, 2019).
(141) Allport, D. C.; Gilbert, D. S.; Outterside, S. M. *MDI and TDI: Safety, Health and the Environment: A Source Book and Practical Guide*; John Wiley & Sons, 2003.
(142) Broderius, S.; Kahl, M. Acute Toxicity of Organic Chemical Mixtures to the Fathead Minnow. *Aquatic Toxicology* **1985**, *6* (4), 307–322. https://doi.org/10.1016/0166-445X(85)90026-8.

(143) Hermens, J.; Canton, H.; Steyger, N.; Wegman, R. Joint Effects of a Mixture of 14 Chemicals on Mortality and Inhibition of Reproduction of Daphnia Magna. *Aquatic Toxicology* **1984**, *5* (4), 315–322. https://doi.org/10.1016/0166-445X(84)90012-2.
(144) Niederlehner, B. R.; Cairns, J.; Smith, E. P. Modeling Acute and Chronic Toxicity of Nonpolar Narcotic Chemicals and Mixtures ToCeriodaphnia Dubia. *Ecotoxicology and Environmental Safety* **1998**, *39* (2), 136–146. https://doi.org/10.1006/eesa.1997.1621.
(145) De Wolf, W.; Canton, J. H.; Deneer, J. W.; Wegman, R. C. C.; Hermens, J. L. M. Quantitative Structure-Activity Relationships and Mixture-Toxicity Studies of Alcohols and Chlorohydrocarbons: Reproducibility of Effects on Growth and Reproduction of Daphnia Magna. *Aquatic Toxicology* **1988**, *12* (1), 39–49. https://doi.org/10.1016/0166-445X(88)90018-5.
(146) Pennington, D. W.; Potting, J.; Finnveden, G.; Lindeijer, E.; Jolliet, O.; Rydberg, T.; Rebitzer, G. Life Cycle Assessment Part 2: Current Impact Assessment Practice. *Environment International* **2004**, *30* (5), 721–739. https://doi.org/10.1016/j.envint.2003.12.009.
(147) Hauschild, M. Z.; Huijbregts, M.; Jolliet, O.; MacLeod, M.; Margni, M.; van de Meent, D.; Rosenbaum, R. K.; McKone, T. E. *Building a Model Based on Scientific Consensus for Life Cycle Impact Assessment of Chemicals: The Search for Harmony and Parsimony*; ACS Publications, 2008.
(148) McKone, T. E.; Kyle, A. D.; Jolliet, O.; Olsen, S. I.; Hauschild, M. Dose-Response Modeling for Life Cycle Impact Assessment-Findings of the Portland Review Workshop. *The International Journal of Life Cycle Assessment* **2006**, *11* (2), 137–140.
(149) Hertwich, E. G.; Mateles, S. F.; Pease, W. S.; McKone, T. E. Human Toxicity Potentials for Life-Cycle Assessment and Toxics Release Inventory Risk Screening. *Environmental Toxicology and Chemistry* **2001**, *20* (4), 928–939. https://doi.org/10.1002/etc.5620200431.
(150) Epa, U. Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11. *United States Environmental Protection Agency, Washington, DC, USA* **2012**.
(151) Franco, A.; Trapp, S. Estimation of the Soil–Water Partition Coefficient Normalized to Organic Carbon for Ionizable Organic Chemicals. *Environmental toxicology and chemistry* **2008**, *27* (10), 1995–2004.
(152) Burkhard, L. P. Estimating Dissolved Organic Carbon Partition Coefficients for Nonionic Organic Chemicals. *Environmental Science & Technology* **2000**, *34* (22), 4663–4668.
(153) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol− Water Partition Coefficients by Guiding an Additive Model with Knowledge. *Journal of chemical information and modeling* **2007**, *47* (6), 2140–2148.
(154) Glende, C.; Schmitt, H.; Erdinger, L.; Engelhardt, G.; Boche, G. Transformation of Mutagenic Aromatic Amines into Non-Mutagenic Species by Alkyl Substituents: Part I. Alkylation Ortho to the Amino Function. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* **2001**, *498* (1–2), 19–37. https://doi.org/10.1016/S1383-5718(01)00259-5.

(155) Nikolova-Jeliazkova, N.; Jaworska, J. An Approach to Determining Applicability Domains for QSAR Group Contribution Models: An Analysis of SRC KOWWIN. *ATLA-NOTTINGHAM-* **2005**, *33* (5), 461.
(156) Lloyd, S. M.; Ries, R. Characterizing, Propagating, and Analyzing Uncertainty in Life-Cycle Assessment: A Survey of Quantitative Approaches. *Journal of Industrial Ecology* **2007**, *11* (1), 161–179. https://doi.org/10.1162/jiec.2007.1136.
(157) Qin, Y.; Suh, S. What Distribution Function Do LCIs Follow? *Int J Life Cycle Assess*.
(158) Sills, D. L.; Paramita, V.; Franke, M. J.; Johnson, M. C.; Akabas, T. M.; Greene, C. H.; Tester, J. W. Quantitative Uncertainty Analysis of Life Cycle Assessment for Algal Biofuel Production. *Environmental science & technology* **2012**, *47* (2), 687–694.
(159) Arena, M.; Azzone, G.; Conte, A. A Streamlined LCA Framework to Support Early Decision Making in Vehicle Development. *Journal of Cleaner Production* **2013**, *41*, 105–113. https://doi.org/10.1016/j.jclepro.2012.09.031.
(160) Wernet, G.; Hellweg, S.; Hungerbühler, K. A Tiered Approach to Estimate Inventory Data and Impacts of Chemical Products and Mixtures. *Int J Life Cycle Assess* **2012**, *17* (6), 720–728. https://doi.org/10.1007/s11367-012-0404-0.
(161) Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *WIREs Comput Mol Sci* **2014**, *4* (5), 468–481. https://doi.org/10.1002/wcms.1183.
(162) Song, R.; Keller, A. A.; Suh, S. Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. *Environ. Sci. Technol.* **2017**, *51* (18), 10777–10785. https://doi.org/10.1021/acs.est.7b02862.
(163) Iooss, B.; Lemaître, P. A Review on Global Sensitivity Analysis Methods. In *Uncertainty management in simulation-optimization of complex systems*; Springer, 2015; pp 101–122.
(164) Cucurachi, S.; Borgonovo, E.; Heijungs, R. A Protocol for the Global Sensitivity Analysis of Impact Assessment Models in Life Cycle Assessment. *Risk Analysis* **2016**, *36* (2), 357–377. https://doi.org/10.1111/risa.12443.
(165) Wei, P.; Lu, Z.; Yuan, X. Monte Carlo Simulation for Moment-Independent Sensitivity Analysis. *Rel. Eng. & Sys. Safety* **2013**, *110*, 60–67. https://doi.org/10.1016/j.ress.2012.09.005.
(166) Pianosi, F.; Wagener, T. A Simple and Efficient Method for Global Sensitivity Analysis Based on Cumulative Distribution Functions. *Environmental Modelling & Software* **2015**, *67*, 1–11. https://doi.org/10.1016/j.envsoft.2015.01.004.
(167) Borgonovo, E. A New Uncertainty Importance Measure. *Reliability Engineering & System Safety* **2007**, *92* (6), 771–784. https://doi.org/10.1016/j.ress.2006.04.015.
(168) US EPA, O. EPI Suite™-Estimation Program Interface https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface (accessed Dec 22, 2017).
(169) Chemistry Dashboard | Home https://comptox.epa.gov/dashboard (accessed Dec 26, 2018).
(170) The PubChem Project https://pubchem.ncbi.nlm.nih.gov/ (accessed Dec 26, 2018).
(171) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *Journal of machine learning research* **2011**, *12* (Oct), 2825–2830.

(172) Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A. Quantitative Nanostructure−Activity Relationship Modeling. *ACS Nano* **2010**, *4* (10), 5703–5712. https://doi.org/10.1021/nn1013484.
(173) Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer series in statistics New York, NY, USA:, 2001; Vol. 1.
(174) Lozano-Perez, T. *Autonomous Robot Vehicles*; Springer Science & Business Media, 2012.
(175) Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. A. Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In *AAAI*; 2017; Vol. 4, p 12.
(176) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems*; 2012; pp 1097–1105.
(177) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263–274. https://doi.org/10.1021/ci500747n.
(178) Prettenhofer, P.; Louppe, G. Gradient Boosted Regression Trees in Scikit-Learn. **2014**.
(179) Lordo, R. A. *Learning from Data: Concepts, Theory, and Methods*; Taylor & Francis, 2001.
(180) Bare, J.; Young, D.; QAM, S.; Hopton, M.; Chief, S. A. B. Tool for the Reduction and Assessment of Chemical and Other Environmental Impacts (TRACI). **2012**.
(181) 2.10.2 Direct Global Warming Potentials - AR4 WGI Chapter 2: Changes in Atmospheric Constituents and in Radiative Forcing https://www.ipcc.ch/publications_and_data/ar4/wg1/en/ch2s2-10-2.html (accessed May 12, 2016).
(182) Goedkoop, M.; Spriensma, R. *The Eco-Indicator99: A Damage Oriented Method for Life Cycle Impact Assessment: Methodology Report*; 2001; pp 1–144.
(183) Morton, M. G.; F. L. Mayer, J.; Dickson, K. L.; Waller, W. T.; Moore, J. C. Acute and Chronic Toxicity of Azinphos-Methyl to Two Estuarine Species, Mysidopsis Bahia and Cyprinodon Variegatus. *Arch. Environ. Contam. Toxicol.* **1997**, *32* (4), 436–441. https://doi.org/10.1007/s002449900210.
(184) Goodman, L. R.; Cripe, G. M.; Moody, P. H.; Halsell, D. G. Acute Toxicity of Malathion, Tetrabromobisphenol-A, and Tributyltin Chloride to Mysids (Mysidopsis Bahia) of Three Ages. *Bull. Environ. Contam. Toxicol.* **1988**, *41* (4–6), 746–753. https://doi.org/10.1007/BF02021028.
(185) Hirano, M.; Ishibashi, H.; Matsumura, N.; Nagao, Y.; Watanabe, N.; Watanabe, A.; Onikura, N.; Kishi, K.; Arizono, K. Acute Toxicity Responses of Two Crustaceans, *Americamysis Bahia* and *Daphnia Magna*, to Endocrine Disrupters. *Journal of Health Science* **2004**, *50* (1), 97–100. https://doi.org/10.1248/jhs.50.97.
(186) Alexander, H. C.; Dill, D. C.; Smith, L. W.; Guiney, P. D.; Dorn, P. Bisphenol a: Acute Aquatic Toxicity. *Environmental Toxicology and Chemistry* **1988**, *7* (1), 19–26. https://doi.org/10.1002/etc.5620070104.
(187) DeLorenzo, M. E.; Key, P. B.; Chung, K. W.; Sapozhnikova, Y.; Fulton, M. H. Comparative Toxicity of Pyrethroid Insecticides to Two Estuarine Crustacean

Species, Americamysis Bahia and Palaemonetes Pugio. *Environ. Toxicol* **2014**, *29* (10), 1099–1106. https://doi.org/10.1002/tox.21840.
(188) Neuparth, T.; Moreira, S.; Santos, M. M.; Reis-Henriques, M. A. Hazardous and Noxious Substances (HNS) in the Marine Environment: Prioritizing HNS That Pose Major Risk in a European Context. *Marine Pollution Bulletin* **2011**, *62* (1), 21–28. https://doi.org/10.1016/j.marpolbul.2010.09.016.
(189) Nimmo, D. R.; Hamaker, T. L.; Matthews, E.; Young, W. T. The Long-Term Effects of Suspended Particulates on Survival and Reproduction of the Mysid Shrimp, Mysidopsis Bahia, in the Laboratory. **1982**.
(190) Buccafusco, R. J.; Ells, S. J.; LeBlanc, G. A. Acute Toxicity of Priority Pollutants to Bluegill (Lepomis Macrochirus). *Bulletin of Environmental Contamination and Toxicology* **1981**, *26* (1), 446–452.