Title
Rapid Life-Cycle Impact Screening Using Artificial Neural Networks

Journal
Environmental Science and Technology, 51(18)

ISSN
0013-936X

Authors
Song, Runsheng
Keller, Arturo A
Suh, Sangwon

Publication Date
2017-09-19

DOI
10.1021/acs.est.7b02862

Peer reviewed

Environmental Science & Technology

Article

pubs.acs.org/est

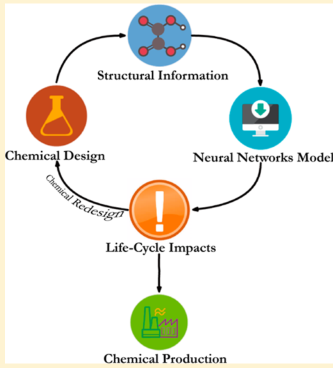
Rapid Life-Cycle Impact Screening Using Artificial Neural Networks

Runsheng Song, Arturo A. Keller, and Sangwon Suh*

Bren School of Environmental Science and Management, University of California, Santa Barbara, California 93106, United States

Supporting Information

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, we developed deep artificial neural network (ANN) models to estimate life-cycle impacts of chemicals. Using molecular structure information, we 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. We 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 values 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.



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 the manufactures to evaluate life-cycle impacts when assessing the alternatives of the priority chemical–application combinations identified.[1] Life-cycle assessment (LCA), among other methods, has been widely used for assessing chemical alternatives.[2−4]

However, in the past, the pace at which LCAs are conducted could not keep up with the speed of new chemical development. According to the Chemical Abstracts Service (CAS), over 100 million unique substances are already registered, and about 15 000 new chemicals are newly added to the list every day.[5] The candidate chemical list of SCP alone contains over a thousand chemicals.[6] Furthermore, the details of new and emerging chemical synthesis are some of the best-protected proprietary information that is rarely disclosed to LCA practitioners, limiting our understanding of their impacts.[7]

Streamlined LCA approaches have been developed and tested to overcome this challenge.[8−11] Such approaches help screen the life-cycle impacts of chemicals without requiring extensive data.[12] Among others, the use of proxy data and regression models are two of the most common approaches to address the data deficiencies in LCA.[13−16] For example, proxy data were used to fill in the data gaps on biobased products,[13] and linear regression models were used to approximate the carbon dioxide emissions from power plants.[15] The level of uncertainty introduced by these approaches may vary widely.[13,17,18]

Another approach to the data gap challenge is the use of machine learning techniques, in which molecular-structure models (MSMs) are used to estimate 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.[19−21] The inherent relationships between molecular structures and potential life cycle impacts of chemical enables MSMs-based estimation of chemical life-cycle impacts.[22] For example, chemicals with long chains, such as polymers, usually require multiple synthesis steps to bond small molecules together requiring more energy and CO_2 emissions throughout the life cycle.[23] 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.[24] Although in some cases, such relationships are not intuitive or obvious to humans, a well-trained MSMs may be able to reveal them.[22]

Wernet and colleagues, for example, applied artificial neural networks (ANN), one of the approaches in MSMs, with one hidden layer 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

Received: June 4, 2017
Revised: August 4, 2017
Accepted: August 15, 2017
Published: August 15, 2017

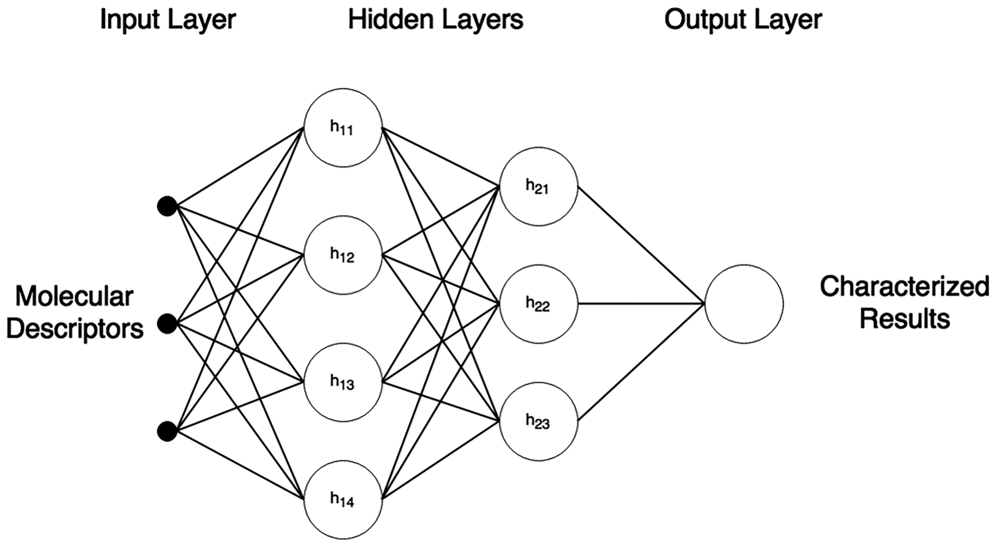


Figure 1. Conceptual diagram for a fully connected ANN model with two hidden layers. The solid lines between layers represent weights that are used in the approximation functions. The value in each node in the hidden and output layers is the sum of the values in the previous layer multiplied by the corresponding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

as input to the models.[26] 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 layers.

In this study, we 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. The training, validation and testing techniques employed in our model are also widely regarded as the state-of-the-art in MSM.[25,26] Furthermore, we also characterized the confidence level of the ANN model outputs using the concept of Applicability Domain (AD), applied for the first time in the context of predictive LCA.[27,28]

This paper is organized as follows: the Materials and Methods section presents the ANN model and the organization of the data used; the Results and Discussion section discusses 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 discussed at the end of this paper.

MATERIALS AND METHODS

Artificial Neural Networks. ANN is a nonlinear, universal approximation model that usually has greater predictive power compared to linear regression, and it also displays significant adaptability for various tasks.[29−31] 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,[32] 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. We then minimize the cost function, which measures the difference between predicted and observed values using the training data set, 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.[33,34]

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 the 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.[35] The ANN models in this study were developed using the Google Tensorflow framework in Python 2.7 under the Ubuntu 16.04 LTS system.[36]

Many successful ANNs utilize large-scale data sets. The Deep Convolutional Neural Network, for example, uses the ImageNet that contains over 10 million URLs of images.[55] However, studies also showed that simpler ANN models based on smaller training data sets can still provide meaningful results.[14,21,26,30] Given the limited availability of LCI data sets for training, we aimed at developing a simpler ANN model.

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, we collected 166 unit process data sets for organic chemicals from the Ecoinvent v3.01 life-cycle inventory (LCI) database.[37] These chemicals were split into three groups for model development, optimization and reporting: training, validation and testing.

We selected three midpoint impact categories: cumulative energy demand (CED),[38] global warming (IPCC 2007, 100a),[39] acidification (TRACI 2.0);[40] and three end point impact categories: eco-indicator 99 (I,I, total) (EI99),[41] ecosystem quality (Impact 2002+),[42] and human health

DOI: 10.1021/acs.est.7b02862
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

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^{-4}	9.8×10^{-5}
standard deviation	41.3	1.0	10.2	0.4	5.1×10^{-4}	9.6×10^{-5}
minimum	19.9	0.1	0.0001	0.01	4.8×10^{-5}	1.3×10^{-6}
median	85.2	1.0	3.2	0.3	4.3×10^{-4}	6.6×10^{-5}
maximum	288.1	6.8	107.9	2.6	3.3×10^{-3}	4.9×10^{-4}

(Impact 2002+).[42] Detailed explanations of these impact categories can be found in the Supporting Information. These six impact categories were chosen to test the model's ability to capture various aspects of chemicals' environmental impact.

Molecular descriptors are a critical component of the training data for our model. They are widely used in computational chemistry and the QSAR field to describe molecular structure.[43] Common descriptors are, for example, molecular weight, number of aromatic rings, number of functional groups and number of halogen atoms.[44] We used the software, *Dragon* 7 to calculate the molecular descriptors for the chemicals in this study.[45] *Dragon* 7 is able to calculate about 4,000 molecular descriptors for each chemical,[46] 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.[47] 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.[48] 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.[49−51] The variables projected after PCA lose the physical meaning of the original molecular descriptors, but they do preserve most of the variances in the original data set. Filter-based feature selection is another method, which removes the descriptors with low variance and high mutual correlation. In filter-based methods, remaining descriptors will preserve the physical meaning of the original descriptors; however, the removed descriptors may contain useful information for the prediction. 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 subset with the best model performance. Due to the high computational cost and the risk of overfitting, however, wrapper-based feature selection method was not chosen for this study.[52]

In this study, we ran and compared the performances 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 variances in the original data set. 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 by feature selection or PCA was normalized by calculating the z-score of them, as shown in eq 1, to have zero mean and unit variance:[53]

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

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.[29] Tuning each hyper-parameter is very time-consuming and, in many cases, unnecessary. In our study, we optimized the number of hidden layers, as well as the number of hidden neurons in each hidden layer using the validation and test data sets. This ensured that the best model structure was used and that the model performance was not affected by the selection of the validation data set.[54]

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 data set used in this study is presented in Table S1.

Model Applicability Domain. Supervised-learning models make predictions based on what the models learn from the training data.[35] 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.

Various AD measurement methods are available and discussed in the QSAR literature.[56−58] Based on the chemical LCI data collected in our study, we applied the Euclidean distance-based AD measurement method.[56] Other AD measurement methods, such as the probability density approaches, were not applicable to the data we collected in this study.[57] The Euclidean distance-based method measures the Euclidean distance in the descriptors' space from the query chemical to the mean of the training data set (namely, the training data centroid). This distance is defined as:

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

where D is the distance between the query chemical X and the training data centroid u; and 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 data set to the centroid of the training data is smaller than a precalculated cutoff threshold. In many QSAR studies, this cutoff threshold is chosen subjectively

DOI: 10.1021/acs.est.7b02862
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

by expert judgements.[57] In our study, we selected the threshold in such a way that the difference between the average prediction error among the data points in the validation data set within the AD and that among the data points outside is the largest. We then applied the selected cutoff threshold to the testing data set.

RESULTS AND DISCUSSION

Chemical Used for Model Development. The chemical data set we 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 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 (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, we transformed the characterized results of global warming, human health, and ecosystem quality models to a log-scale before training.

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 data set, based on: (1) all the descriptors generated by Dragon 7 (3839 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). We examined each of the three cases with 1, 2, or 3 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 data set without the testing data set.

As shown in Figure 2, the ANN models for acidification using all descriptors showed the lowest R^2 values. Although the discrepancy is not significant, the descriptors extracted using PCA resulted in a better performance in 8 out of 12 models as compared with 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 data sets comprising all, feature-selection-based, and PCA-based descriptors, respectively. The same analysis for the ANN models of other impact categories can be found from Tables S4–9. 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 data set. As a result, we employed PCA as the approach to reduce the dimensions in the input data and to improve the ANN's performance.

Figure 3 shows the results of optimization for the CED and EI99 models. The models were developed with the descriptors

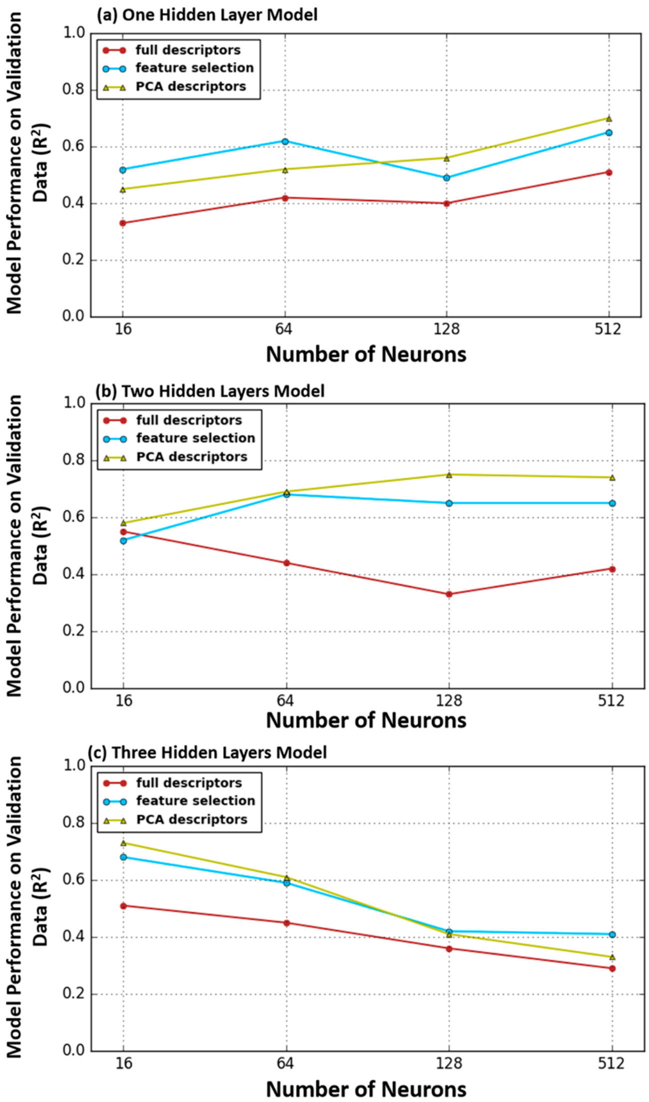


Figure 2. Performance (R^2) of the acidification model developed with (1) all molecular descriptors set (red); (2) molecular descriptors after feature selection (blue); and (3) molecular descriptors after PCA (yellow). The performances are the results using the validation data set without the testing data set. The same analysis for the other models can be found from Tables S4−S9.

extracted by PCA and the performance was measured using the validation data set. 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 data set decreased and overfitting occurred as we 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 did 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 we could collect.

Based on the validation results, optimized model structure is presented in Table 2. The human health model requires the

DOI: 10.1021/acs.est.7b02862
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

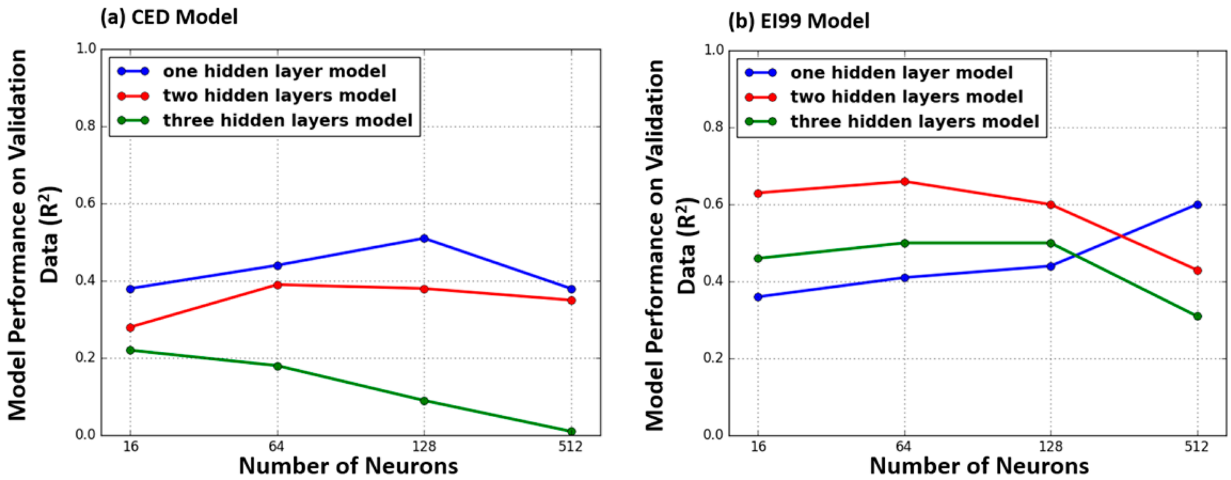


Figure 3. Model performance (R^2) using the validation data set for (a) the CED model and (b) the EI99 model with one, two, and three hidden layer(s) and 16, 64, 128, and 512 hidden neurons embedded in each layer. Descriptors selected using PCA were considered as the input.

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[a]	1	128
acidification	2	128
EI99[b]	2	64
global warming	2	16
human health	3	64
ecosystem quality	2	128

[a]Cumulative energy demand. [b]EI99: eco-indicator 99.

highest complexity (3 hidden layers with 64 hidden neurons in each layer) among all models. The detailed hyper-parameters, such as the learning rate, activation function, and training epoch, can be found in Table S10.

Model Performance. A total of 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 data sets are reported (R^2 and mean relative error (MRE)) in Figure 4 and Table 3. Each panel in Figure 4 shows the model performance for the corresponding impact category. Circles represent the performance on the training data set, squares the performance on the validation data set and triangles the model performance on the testing data set. 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 values of 0.73, 0.87 and 0.71, respectively. The CED and ecosystem quality models showed lower performance, with R^2 values of 0.45 and 0.48 on the testing data set, respectively. The performance of the global warming model was the lowest of all. Even though the R^2 on the testing data set was 0.48, the training and validation accuracy of the global warming model 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 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. Compared to the model presented in Wernet et al. (2008), our models show a significant improvement on EI99 (0.87 versus 0.67, in R^2), while the R^2 values for CED and global warming results are comparable between the two. However, it is notable that a direct comparison of the model performance between the two ANN models based only on R^2 values is difficult because the chemicals used as the testing data are different.

Model Applicability Domain Analysis. The MRE of both the validation and testing data sets that fall within and outside of the AD in each model are presented in Table 4. The testing data set 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.

Case Study. We selected two chemicals, acetic anhydride and hexafluoroethane (HFE), from the testing data set for a case study to demonstrate how our models work. 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 shown 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 of the 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

DOI: 10.1021/acs.est.7b02862
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

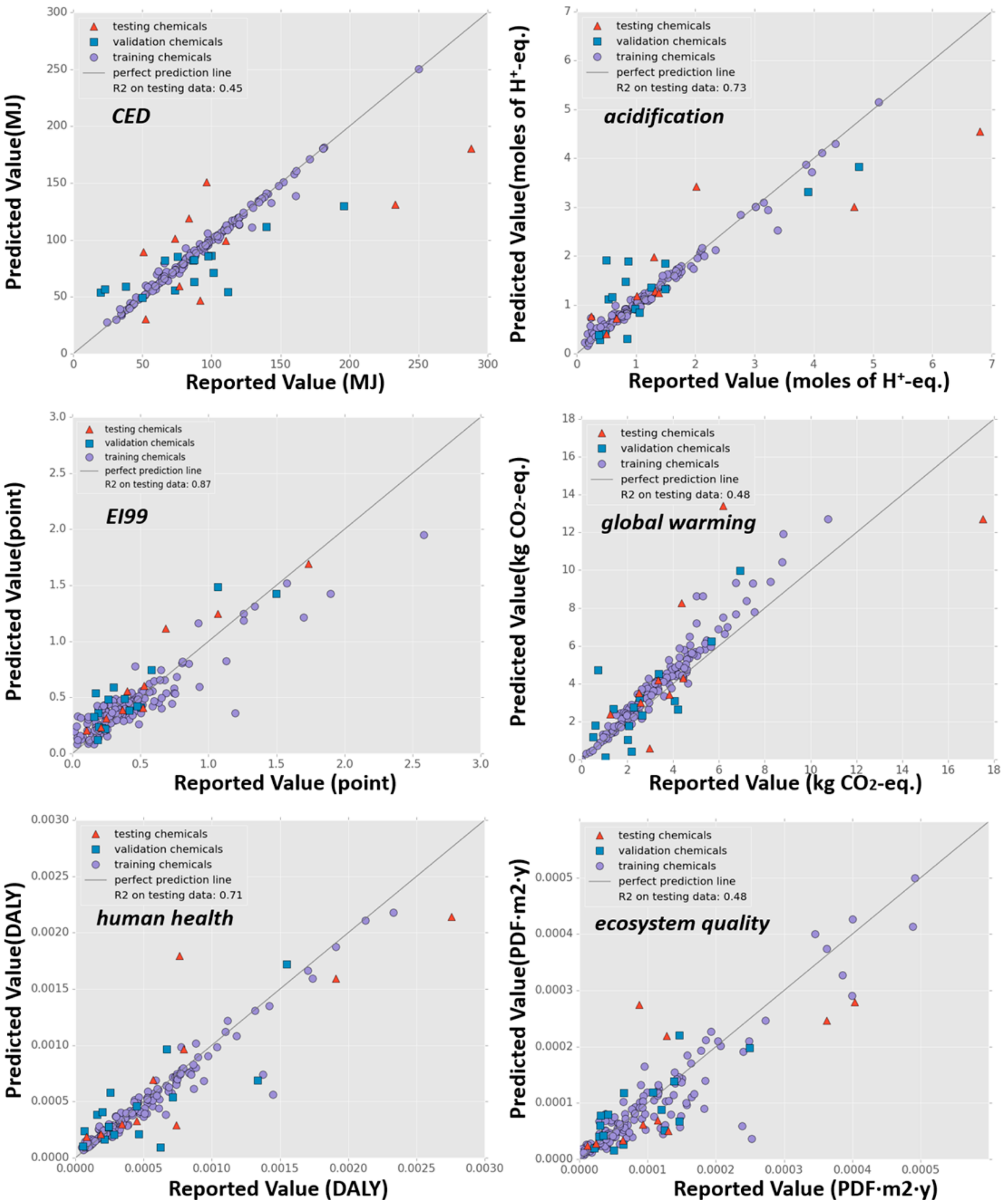


Figure 4. Model performance considering the training, validation, and testing data sets. The training data set was used to develop each model. The validation data set was used to optimize the model structure, and the testing data set was used to report the model performance.

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

		CED[a]	acidification	EI99[b]	global warming	human health	ecosystem quality
training data set	R^2	0.98	0.97	0.82	0.31	0.94	0.84
	MRE	3%	14%	55%	20%	15%	47%
validation data set	R^2	0.52	0.75	0.72	0.21	0.58	0.48
	MRE	40%	56%	50%	88%	68%	52%
testing data set	R^2	0.45	0.73	0.87	0.48	0.71	0.48
	MRE	40%	46%	30%	50%	46%	65%

[a]Cumulative energy demand. [b]EI99: eco-indicator 99.

DOI: 10.1021/acs.est.7b02862
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 4. Mean Relative Error (MRE) of Chemicals Inside and Outside of the Measured AD on Both Validation and Testing Dataset for Each Model[a]

	validation data set		testing data set	
	MRE within AD	MRE outside AD	MRE within AD	MRE outside AD
CED[b]	18%	47%	30%	44%
acidification	32%	150%	26%	76%
EI99[c]	36%	107%	21%	43%
global warming	25%	92%	65%	50%
human health	62%	180%	75%	111%
ecosystem quality	41%	104%	40%	63%

[a]The AD was measured on the validation dataset. [b]Cumulative energy demand. [c]EI99: eco-indicator 99.

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 the global warming model, and HFE is located outside of every model's AD.

Limitations and Recommendations. The MSMs we 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 have been used to analyze the contribution of each molecular descriptor, but the prediction accuracy is reported to be low.[26]

Because we 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 data sets are derived using crude assumptions, heuristic rules, and stoichiometric relationships. The outputs of the models using such data as the training data set would provide comparable results with the existing data sets because they cannot overcome the limitations of the data sets.

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 nonparametric probability density distribution method, can be considered as a means to improve the AD measurement when training data are normally distributed.[57]

Future research may 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. Future studies should consider using cross-validation techniques to avoid the potential bias in the selection of training data, especially when the model uses a single layer. Most of all, improving the availability of reliable and harmonized LCI data would be crucial to develop reliable ANN models for LCA. A larger LCI database with diverse chemical types can benefit from the use of more-complex ANN model structures, which may help improve the performance of predictive LCA.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.7b02862.

Additional details on training chemicals, molecular descriptors used in this study, impact categories, model optimization and development, and model applicability domain measurements. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +1-805-893-7185; fax: +1-805-893-7612; e-mail: suh@bren.ucsb.edu.

ORCID

Arturo A. Keller: 0000-0002-7638-662X

Sangwon Suh: 0000-0001-8290-6276

Notes

The authors declare no competing financial interest.

Table 5. Model Estimation Results of Acetic Anhydride and HFE for the Six Selected Impact Categories in This Study, along with the AD Analysis for These Two Chemicals[a]

	acetic anhydride	hexafluoroethane
within AD?	yes[b]	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%)[c]	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%)

[a]The numbers shows reported values and the values in the parentheses are values estimated by the model and the absolute value of relative error. [b]Excluding global warming model. [c]Out of AD.

DOI: 10.1021/acs.est.7b02862
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

■ ACKNOWLEDGMENTS

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 the U.S. EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. The U.S. EPA does not endorse any products or commercial services mentioned in this publication.

■ REFERENCES

(1) California Department of Toxic Substances Control. Alternatives Analysis Guide. https://www.dtsc.ca.gov/SCP/AlternativesAnalysisGuidance.cfm (accessed Nov 19, 2015).

(2) 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.

(3) Eckelman, M. J. Life Cycle Inherent Toxicity: A Novel LCA-based Algorithm for Evaluating Chemical Synthesis Pathways. *Green Chem.* **2016**, *18*, 3257.

(4) Anastas, P. T.; Lankey, R. L. Life cycle assessment and green chemistry: the yin and yang of industrial ecology. *Green Chem.* **2000**, *2*, 289−295.

(5) Chemical Abstracts Service. CAS Assigns the 100 Millionth CAS Registry Number® to a Substance Designed to Treat Acute Myeloid Leukemia. https://www.cas.org/news/media-releases/100-millionth-substance (accessed Nov 20, 2015).

(6) Chemical Abstracts Service. Candidate Chemical List. https://www.dtsc.ca.gov/SCP/CandidateChemicals.cfm (accessed Jan 22, 2016).

(7) Hischier, R.; Hellweg, S.; Capello, C.; Primas, A. Establishing Life Cycle Inventories of Chemicals Based on Differing Data Availability (9 pp). *Int. J. Life Cycle Assess.* **2005**, *10* (1), 59−67.

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) Hanes, R.; Bakshi, B. R.; Goel, P. K. The use of regression in streamlined life cycle assessment. In *Proceedings of the International Symposium on Sustainable Systems and Technologies*; ISSST: Tempe, AZ, 2013.

(18) Niero, M.; Di Felice, F.; 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.

(19) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Neural networks in building QSAR models. *Methods Mol. Biol. (N. Y., NY, U. S.)* **2009**, *458*, 133−154.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) von der Assen, N.; 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.

(25) 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.

(26) 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.

(27) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26* (5), 694−701.

(28) OECD. Validation of (Q)SAR Models - OECD. http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed April 26, 2016).

(29) Kalogirou, S. A. Artificial neural networks in renewable energy systems applications: a review. *Renewable Sustainable Energy Rev.* **2001**, *5* (4), 373−401.

(30) 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.

(31) Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **1989**, *2* (5), 359−366.

(32) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, *61*, 85−117.

(33) Hanrahan, G. *Artificial neural networks in biological and environmental analysis*; CRC Press: Boca Raton, FL, 2011.

(34) 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.

(35) Hanrahan, G. *Artificial Neural Networks in Biological and Environmental Analysis*; CRC Press: Boca Raton, FL, 2011.

(36) 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. 2016,

DOI: 10.1021/acs.est.7b02862
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

arXiv:cs/1603.04467. arXiv.org e-Print archive. https://arxiv.org/abs/1603.04467 (accessed Mar 14, 2017).

(37) 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.

(38) 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.

(39) IPCC. *Climate Change 2007: The physical science basis*; Cambridge University Press: Cambridge, UK, 2007.

(40) Bare, J. TRACI 2.0: the tool for the reduction and assessment of chemical and other environmental impacts 2.0. *Clean Technol. Environ. Policy* **2011**, *13* (5), 687−696.

(41) Goedkoop, M.; Spriensma, R. *The Eco-indicator99: a damage oriented method for life cycle impact assessment: methodology report*; PRé: Amersfoort, The Netherlands, 2001; pp 1−144.

(42) 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. Life Cycle Assess.* **2003**, *8* (6), 324−330.

(43) Karelson, M. *Molecular descriptors in QSAR/QSPR*; Wiley-Interscience: Hoboken, NJ, 2000.

(44) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18* (4−5), 464−477.

(45) Kode - Chemoinformatics https://chm.kode-solutions.net/products_dragon.php (accessed Apr 26, 2017).

(46) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.

(47) 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.

(48) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.

(49) 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.

(50) 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.

(51) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313* (5786), 504−507.

(52) Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23* (19), 2507−2517.

(53) 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.

(54) Mosteller, F.; Tukey, J. W. Data Analysis, Including Statistics. In *Handbook of Social Psychology*, vol 2; Lindzey, G., Aronson, E., Eds.; Addison-Wesley: Boston, MA, 1968.

(55) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, 1097−1105.

(56) 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.

(57) 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.

(58) 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.

DOI: 10.1021/acs.est.7b02862
Environ. Sci. Technol. XXXX, XXX, XXX−XXX