Environmental Science & Technology

pubs.acs.org/est

Article

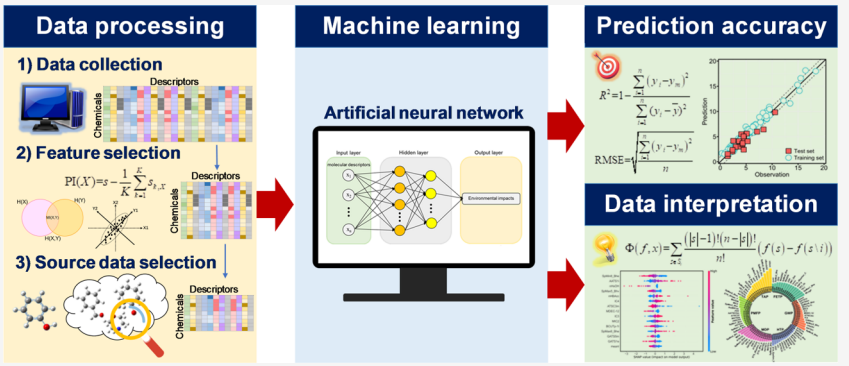
Improved Machine Learning Models by Data Processing for Predicting Life-Cycle Environmental Impacts of Chemicals

Ye Sun, Xiuheng Wang, Nanqi Ren, Yanbiao Liu, and Shijie You*

Cite This: *Environ. Sci. Technol.* 2023, 57, 3434−3444

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Machine learning (ML) provides an efficient manner for rapid prediction of the life-cycle environmental impacts of chemicals, but challenges remain due to low prediction accuracy and poor interpretability of the models. To address these issues, we focused on data processing by using a mutual information-permutation importance (MI-PI) feature selection method to filter out irrelevant molecular descriptors from the input data, which improved the model interpretability by preserving the physicochemical meanings of original molecular descriptors without generation of new variables. We also applied a weighted Euclidean distance method to mine the data most relevant to the predicted targets by quantifying the contribution of each feature, thereby the prediction accuracy was improved. On the basis of above data processing, we developed artificial neural network (ANN) models for predicting the life-cycle environmental impacts of chemicals with R^2 values of 0.81, 0.81, 0.84, 0.75, 0.73, and 0.86 for global warming, human health, metal depletion, freshwater ecotoxicity, particulate matter formation, and terrestrial acidification, respectively. The ML models were interpreted using the Shapley additive explanation method by quantifying the contribution of each input molecular descriptor to environmental impact categories. This work suggests that the combination of feature selection by MI-PI and source data selection based on weighted Euclidean distance has a promising potential to improve the accuracy and interpretability of the models for predicting the life-cycle environmental impacts of chemicals.

KEYWORDS: *life cycle assessment (LCA), machine learning, data processing, feature selection, weighted Euclidean distance*

1. INTRODUCTION

Recently, the concept of green chemistry has been widely acknowledged and applied in the field of chemical industry due to the rising environmental concerns.[1,2] Chemical regulations should focus on the product's life-cycle aspects rather than end-of-pipe production.[3,4] Life cycle assessment (LCA) provides a useful tool to evaluate and compile the potential environmental impacts of different products and processes throughout the entire life cycle.[5−8] Life cycle inventory (LCI) serves as the foundation of LCA to summarize the total consumption of resources, waste flows, and emissions, so that the life cycle impact assessment (LCIA) can be implemented to quantify the environmental impacts by multiplying LCI with corresponding impact indicators.[9−11] However, only a fraction of chemicals that are synthesized or commercially used are available in the existing databases.[12] Challenges remain for LCA due to limitations of time, proprietary information, and expertise requisite for compiling a complete LCI data set.[3,4]

Much effort has emerged to provide reliable databases accessible for LCA practitioners without the need to collect data from the industry or conduct all the detailed process calculations and simulations.[13−15] Existing methods to estimate missing environmental impact data of chemicals include process simulation, engineered process calculations, molecular structure models, mixed-integer programming methods, stoichiometric

Received: July 11, 2022
Revised: December 5, 2022
Accepted: December 5, 2022
Published: December 20, 2022

methods, and the use of proxies and omission.[12,16,17] For example, several studies have focused on toxicity prediction of chemicals by using computational methods such as uncertainty factor models, dose−response, and time-response models.[18−21] Among them, models derived from machine learning (ML) without requirement for extensive data can circumvent the uncertainties introduced by conventional process-based models.[13,22−27] The toxicity and potential environmental impacts can be estimated according to the chemicals' molecular structures by using ML models.[28−32]

Currently, artificial neural network (ANN) has been developed as an initial screening tool for estimating life-cycle environmental impacts of chemicals for certain impact categories in the absence of more reliable information on the molecular structures, where large numbers of molecular descriptors generated by software are used as input data.[3,13] Within this context, data processing is of particular importance to improve the prediction accuracy and reduce the computational complexity before the models are developed.[33,34] To enable more efficient training and avoid possible overfitting, it is highly desirable to reduce the number of dimensions by extracting a more informative subset of descriptors. Song *et al.* reported that the feature extraction (*i.e.*, principal component analysis) performed better in selecting molecular descriptors compared with feature selection such as filter-/wrapper-based methods.[4] Nevertheless, challenges remain for feature extraction because of the needs for transformation of the original data to new variables with strong pattern recognition ability, leading to the loss of the physical meaning of the original molecular descriptors.[35] In comparison, better interpretability can be achieved for feature selection methods due to no generation of any new variable.[33,36] Generally, filter-/wrapper-based methods are the two most commonly adopted feature selection techniques. The filter method is computationally efficient because it is independent of learning algorithms and usually focuses on inventing metrics to depict the relationship of each subset of input features with the output.[4,37,38] Conversely, the wrapper method takes prediction accuracy as the quality criterion for evaluating appropriateness of the feature subsets. Hence, the hybrid filter-wrapper method may combine their inherent advantages, thereby most of the irrelevant and redundant features can be filtered out.[39,40] The power of hybrid method is demonstrated by Hu *et al.*, who reported the application of feature selection in short-term electricity load forecasting.[38] However, less attention has been paid to the hybrid filter-wrapper method in data processing of the ML models for predicting life-cycle impacts of chemicals.

After feature selection, mining the most relevant source data relevant to the prediction target constitutes also an essential step to prediction accuracy.[41] Zhang *et al.* built ML models to predict adsorption of organic compounds, and they demonstrated the cosine similarity to be a suitable metric for source data processing in comparison with Euclidean and City-block methods.[34] However, these methods may be subject to relatively large errors because they do not consider the contribution similarity of each feature.[42] This shortage can be alleviated by weighted Euclidean distance by virtue of the capability of weights to quantify contribution of each variable to the similarity metric.[43] Compared with traditional Euclidean and Manhattan distance, the weighted method is more favorable for measuring the physical distances with nonlinear nature and thus lowering the errors.[42,44] Up to the present, however, the weighted Euclidean distance has not yet been explored as a similarity metric method in source data processing.

In this study, we make the first attempt to perform data processing by a feature selection method and weighted Euclidean distance, both of which are important in order to increase prediction accuracy and interpretability of the ANN-based ML models for predicting the life-cycle impacts of chemicals. First, the original data set was collected and preprocessed, followed by comparison of four ML algorithms. Second, informative descriptor subsets were screened from the preprocessed data set by using a hybrid filter-wrapper method (*i.e.*, mutual information-permutation importance; MI-PI method). Third, to improve the prediction accuracy, the weighted Euclidean distance method was employed to further screen out the source data that are similar as the target prediction chemicals to develop ANN models. Last, the model interpretation was elucidated according to Shapley values and the most relevant molecular descriptors to each of environmental impact categories.

2. MATERIALS AND METHODS

2.1. Data Collection and Preprocessing. Unit process data sets (187) of organic chemicals were collected from the Ecoinvent v3.8 LCI database for development of machine learning models. The Ecoinvent database covered a diverse range of sectors on the global and regional level so that the users were able to gain a deeper understanding of the impacts of the products and services.[45,46] The models were tested by selecting six environmental impact categories including climate change (GWP), particulate matter formation (PMFP), terrestrial acidification (TAP), freshwater ecotoxicity (FETP), human toxicity (HTP), and metal depletion (MDP) based on the ReCiPe Midpoint (H) method. The detailed data of the impact categories were listed in the Supporting Information. Padel-Descriptor software[47] was used to generate the molecular descriptor based on the Simplified Molecular Input Line Entry System (SMILES)[48] of each chemical, generating 1444 descriptors, which are relevant to the nature of chemicals such as the atom, electrotopological state, molecular linear free energy relation, and ring counts. These molecular descriptors were also used as the model input features.

Since the large number of molecular descriptors generated by Padel-Descriptor software led to inefficient training and overfitting, preprocessing was needed for the collected data set. First, descriptors with a variance of zero were removed due to negligible impact to the results of model training. One of any two features with high pairwise correlation (*i.e.*, Pearson correlation coefficient of greater than 0.95) was also removed, and thus 531 features were retained. The z-score Normalization method was used to normalize the features according to eq 1:[49]

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

where Z is the feature after normalization, X the original feature before normalization, μ the mean value of the feature across all the chemicals, and σ the standard deviation of the feature across all the chemicals. The preprocessed data subset was named FULL for further processing.

2.2. Dimensionality Reduction of Data Set. The data subset after preprocessing was characterized by a large number of features (531) and a small number of available samples (187). Prior to model fitting, irrelevant features were filtered out through dimensionality reduction including principal compo-

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444

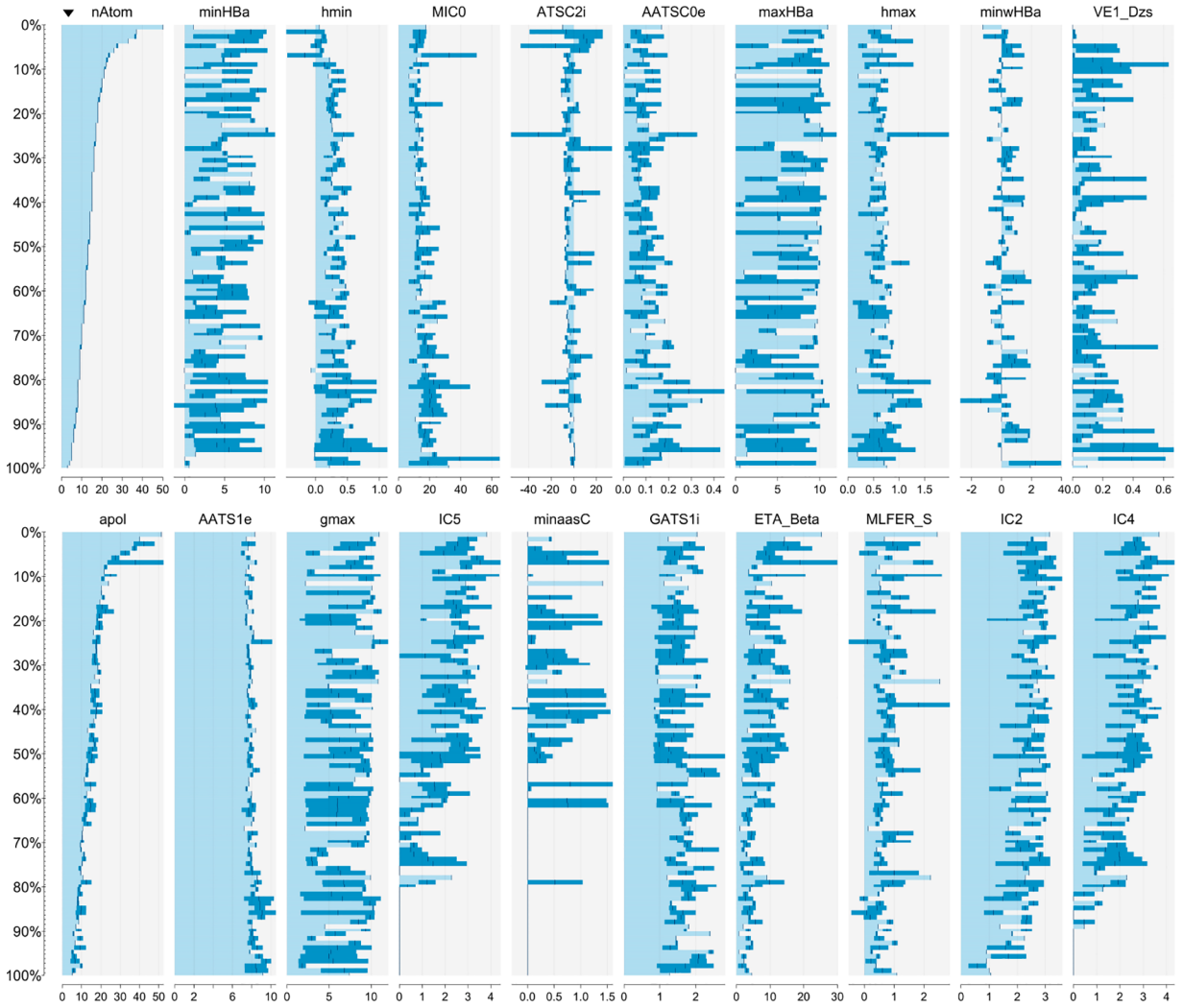


Figure 1. Visualization of data distribution of the descriptors according to the descending order of *n*Atom.

nent analysis (PCA), mutual information (MI) and permutation importance (PI), and the procedures for each method were described in detail in the Supporting Information. To improve the performance of the models, we used a feature selection method based on mutual information and permutation importance (MI-PI) according to the procedures described as follows: (1) The MI method was used first to filter out the redundant and irrelevant features. (2) The features retained by the MI method in the previous step were used to build the models. (3) The features were further selected by using the PI method based on the models built in step (2). (4) The models were rebuilt with the features obtained in step (3).

2.3. Model Construction and Analysis. Four commonly used machine learning algorithms, *i.e.*, random forest (RF), extreme gradient boosting (XGBoost), support vector machine (SVM), and artificial neural network (ANN), were studied and compared toward the prediction of the life-cycle environmental impact of chemicals (Supporting Information). The input data were the preprocessed data set (FULL) containing 531 molecular descriptors, and the output data were the results of an environmental impact category in terms of GWP, HTP, MDP, FETP, PMFP, and TAP. Five different modeling cases were investigated using the following training sets: (1) FULL, all 531 features after data preprocessing; (2) PCA, features extracted by PCA that preserved 95% of the variance in the data set FULL; (3) MI, features selected by MI method; (4) PI, features selected by the PI method; and (5) MI-PI, features selected by the MI-PI method. All the models were programmed by Python 3.7 under the Linux system. The data set was randomly split into two parts, *i.e.*, the training set (90%) and the test set (10%). The GridSearch method with a 10-fold cross validation strategy in the scikit-learn library of Python was used to obtain the optimal hyper-parameters and avoid overfitting of the model.[34,50]

2.4. Model Improvement Based on the Weighted Euclidean Distance. To further improve the performance of the models, the weighted Euclidean distance[43,44] was applied to screen the data subset that was most similar as the target chemical for model training. The weighted Euclidean distance between the target chemical and the training data ($D(p,q)$) was defined as

$$D(p, q) = \sqrt{\sum_{i=1}^{n} w_i^2 (p_i - q_i)^2} \quad (2)$$

where p_i and q_i are the ith molecular descriptors of the target chemical and the training data and w_i the weight of the ith molecular descriptor.

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444

Since the PI method can measure the feature importance under the guidance of model prediction accuracy, the weight of the feature (w_i) was defined as

$$w_i = \frac{PI(X_i)}{\sum_{i=1}^{n} PI(X_i)} \tag{3}$$

where $PI(X_i)$ is the PI value of the ith feature. The detailed description of the method can be found in the Supporting Information (Text S4).

2.5. Model Validation. The performance of the models was evaluated by comparing the coefficient of determination (R^2) and root-mean-square error (RMSE) for the data sets using eqs 4 and 5 upon the test set according to[26]

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_m)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_m)^2}{n}} \tag{5}$$

where y_m is the real value of the environmental impacts, y_i the predictive value of the model output, $\bar{y}$ the mean value of the model output, and n the number of input-output pairs in the data set.

2.6. Mechanistic Interpretation. The SHAP approach was used to identify the importance of the input features and their influence on each of the environmental impacts. The Shapley values were calculated by

$$\Phi(f, x) = \sum_{s \in S_i} \frac{(|s| - 1)!(n - |s|)!}{n!} (f(s) - f(s\backslash i)) \tag{6}$$

where f is the model trained by a set of features x, $\Phi(f,x)$ the Shapley value of feature i, n the number of input features for model f, S_i the set of all possible combinations including feature i, $|s|$ the number of features of one combination in S_i, and $f(s)$ and $f(s\backslash i)$ are the predictions by the model trained on s including i and excluding i, respectively. The feature i with positive Shapley value will push the prediction result in the positive direction during the model prediction, and vice versa. The Shapley values were calculated using the Python implementation with the SHAP package.

3. RESULTS AND DISCUSSION

3.1. Data Set Characterization. The chemical data set collected and pretreated for model training according to the procedures described above could represent a wide range of chemicals such as aromatic compounds, dyes, pharmaceuticals, and halogenated hydrocarbons. The chemicals collected were listed in Table S1, including the SMILES format and environmental impact values for each. The mean, minimum, median, maximum, and standard deviation of each impact category were also calculated (Table S2). Since there was strong variation of the data sets for different impact categories, it was essential to train the models individually (GWP, HTP, MDP, FETP, PMFP, and TAP) and tune the hyper-parameters of each model, respectively. Moreover, to intuitively observe the data set distribution, part of descriptors was selected randomly to illustrate data distribution (Figure 1 and Figure S1), according to the descending order of the number of atoms (nAtom) and GWP, respectively. Since the data set distribution of most descriptors was characterized to be nonlinear and heterogeneous, it appeared unlikely to present the relationship between the molecular descriptors of chemicals and their environmental impacts by general multiple linear regressions. To address this issue, we attempt to use machine learning to deal with this kind of data set on the basis of various machine learning algorithms.[51−54]

3.2. Model Selection and Data Dimensionality Reduction. Four commonly adopted machine learning algorithms, *i.e.*, SVM, RF, XGBoost, and ANN, were investigated, and the results were summarized in Table S4. Specifically, the SVM models exhibited a very poor prediction performance compared with the other three algorithms when the six impact categories were trained, indicated by R^2 = 0.35 and RMSE = 3.34 in the GWP model. This failure should be the consequence of the wide range of data in the training data set, since SVM maximizes the "margin" and therefore depends on the concept of "distance" between different observations.[9] When the value ranges of different descriptors varied considerably, the "distance" became insignificant, which resulted in poor model performance. The RF and XGBoost method yielded the results of the GWP model that were better than SVM (R^2 = 0.56 and RMSE = 0.55). Notably, the RMSE value of 1.89 for the ANN model was lower than that of 2.14 for RF and that of 2.05 for XGBoost, respectively. That is, the ANN model achieved the best performance among all the methods tested in terms of majority impact categories, particularly for TAP with the R^2 value reaching 0.57 and the RMSE value as low as 0.01. Therefore, both the R^2 and RMSE values confirmed the ANN method to be suitable for data set training and further analyses.

Besides the selection of machine learning algorithms, mining valuable data from the abundance of data is also crucial for improving the accuracy of model prediction.[34] After data preprocessing as described in section 2.1, the data subset remained to be characterized by a high dimension containing a large number of features (531) and a small number of available samples (187). To further improve the model prediction accuracy and alleviate overfitting, insignificant features need to be filtered out through dimensionality reduction prior to model fitting. To this end, PCA, MI, PI, and MI-PI methods were studied and compared. Table 1 listed the number of input features after data dimensionality reduction through four different methods. Clearly visible is that dimensionality reduction through MI-PI and PCA methods could greatly reduce the number of input features. Specifically, for the GWP model, the number of input features processed by MI-PI was only 67 compared with the individual MI and PI methods, which was as much as 345 and 186, respectively. For the HTP model, the number of input features processed by MI-PI was reduced to 58 compared with MI (420) and PI (165) methods. Figure 2 shows the performance (R^2) obtained for (a) GWP, (b) HTP, (c) MDP, (d) FETP, (e) PMFP, and (f) TAP ANN models (10-fold ShuffleSplit cross-validation) for dimensionality reduction

Table 1. Number of Input Features after Data Dimensionality Reduction by Four Different Methods

	number of input features of the models					
	GWP	HTP	MDP	FETP	PMFP	TAP
PCA	50	50	50	50	50	50
MI	345	420	404	381	367	333
PI	186	165	189	178	141	153
MI-PI	67	58	83	78	76	86

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444

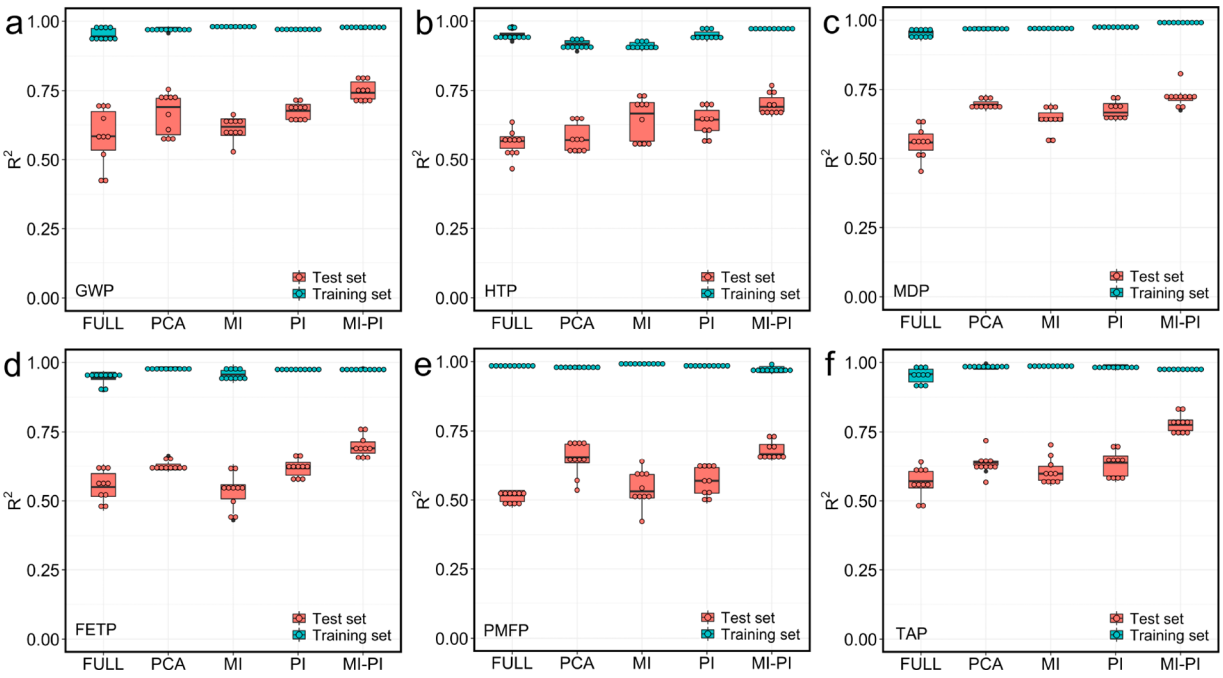


Figure 2. Performance (R^2) of the (a) GWP, (b) HTP, (c) MDP, (d) FETP, (e) PMFP, and (f) TAP ANN models (10-fold ShuffleSplit cross-validation) based on five different data dimensionality reduction approaches.

of the data set. The corresponding RMSEs are summarized in Table S5, and the prediction performance of each model is shown in Figures S2−S7, respectively. For the test set, ANN models trained with the data set of dimensionality reduction by the MI-PI method (*i.e.*, MPN models) were shown to perform the best results toward all the impact categories (maximal R^2 value and minimal RMSE) compared with other methods. For example, the MPN model for GWP had a higher R^2 of 0.75 and a lower RMSE of 1.36, while the model for GWP trained with the data set of dimensionality reduction by the PCA method yielded R^2 of 0.66 and RMSE of 1.38. This suggested that the input data set retained by the MI-PI method could be effective and robust for successfully extracting useful information on the input data set and improving the model prediction accuracy. This also suggested that the overfitting caused by data complexity could be reduced by dimensionality reduction. It is worth noting that this overfitting was more dependent on the quality of data but was irrelevant to the number of descriptors.[55] For example (Table 1), although the number of remaining features of the MI-PI method (67) of the GWP model was larger than that of the PCA method (50), the performance of the model trained by MI-PI features (test set R^2 of 0.75) was better than that of PCA (test set R^2 of 0.66; Figure 2a). The most likely explanation was that the subset of descriptors selected by the MI-PI method contained less irrelevant and redundant information, which should be responsible for alleviating the overfitting of the models.[56]

The most likely reason for superior performance of the MI-PI method should be the complementary feature of MI and PI. In general, MI is a model-free criterion method that can select the input subset and measure the dependency among variables by virtue of high computational efficiency.[37,38] However, it is usually subject to inaccurate prediction due to the lack of guidance for model prediction accuracy.[38] PI can alleviate this shortage by extracting the featuring subset under the guidance of model prediction accuracy, but the permutation of response vector and computation of feature importance by the PI method are quite time-consuming.[36,57,58] This could be illustrated herein by the model for GWP with the running time of only 2.3 s for feature selection by the MI method under the operating environment of Linux Operating System, CPU of Intel Core i7 2.2 GHz, and Memory of 16 GB. For the PI method, the time span as long as 366 s was more than 2 orders of magnitude longer than MI. The combined MI-PI method was expected to reach a compromise between computational efficiency and prediction accuracy, indicated by the running time of only 47 s required for feature selection by the MI-PI method, the value being much shorter than that of individual PI. Moreover, compared with the widely used PCA, the MI-PI method could achieve dimensionality reduction of data set based on preserving the inherent features without loss of model interpretability, making the models more informative and more reliable. On the basis of these results, the MI-PI method was employed for feature selection to develop ANN models.

3.3. Improvement of Model Prediction Accuracy. During the phase of model training, it was found that when the prediction was satisfactory (*e.g.*, R^2 = 0.75 and 0.70 for GWP and HTP model, respectively), the ranges of the feature values for training and test sets were similar (Figures S8a and S9a). However, when the model prediction was poor (R^2 = 0.53 and 0.51 for GWP and HTP model, respectively), the ranges of the feature values for training and test sets were greatly different (Figures S8b and S9b). The apparent meaning of such similarity could be interpreted by the fact that the chemicals with similar descriptor natures (*i.e.*, structural characteristics) might possess similar life-cycle environmental impacts. On the basis of this fact, one expected to screen part of the data with descriptors that were similar to the target predicted chemicals as the training data set for model training rather than all of the collected data. To this end, the weighted Euclidean distance was introduced to measure the similarity of the trained data and target chemicals on a quantitative basis, and the screened data subset containing 60%

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444

of the most relevant data was selected from the original training set for model training. Moreover, commonly adopted similarity metrics, *i.e.*, cosine similarity and Euclidean distance, were also studied and compared, and the results were illustrated in Figure 3 and Table S6. Clearly visible is that the weighted Euclidean distance that measured the similarity under the guidance of the feature's weight was more suitable for our data set, indicated by $R^2 = 0.81$ for GWP compared with $R^2 = 0.75$ for cosine similarity and $R^2 = 0.74$ for Euclidean distance, respectively. Both Euclidean distance and cosine similarity methods were subject to interference with the similarity quantification by the features with low contribution to the models because they did not take feature importance into account. Such shortage could be mitigated by weighted Euclidean distance in favor of its capability to quantify the contribution of each feature to the models based on their weights. The detail description could be found in Text S4. The prediction performances of MPN models and the improved ANN models based on weighted Euclidean distance (MPEN models) were shown in Figure 4 and Figure S10. Clearly visible was that the MPEN models could provide much better predictions than the MPN models for all the target impact categories. For example, $R^2 = 0.81$ was obtained for the MPEN model of GWP, the value being higher than that of the MPN model ($R^2 = 0.75$), accounting for relatively low RMSE of 0.99 for the GWP MPEN model compared with that of the MPN model (1.36). For the HTP MPEN model, the R^2 value also reached a high level of 0.81 with RMSE of 1.11, indicating higher prediction accuracy than the HTP MPN model. Compared with the models reported in previous studies ($R^2 = 0.48$ and 0.71 for GWP and HTP model, respectively),[4] our models showed better prediction accuracy. Meanwhile, the prediction performance also reveals that most of the scatters could fall within the scope of the RMSE interval, which further confirmed accuracy of the models.

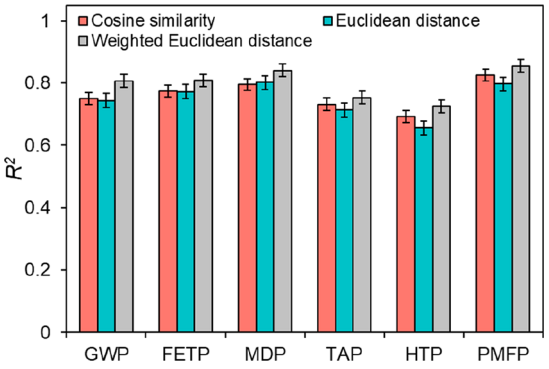


Figure 3. Comparison of the performance obtained for the models based on three approaches to training set selection.

The weighted Euclidean distance was rendered capable of screening out the chemicals that were most relevant to the target chemicals to be predicted.[1] Therefore, the training set containing selected data could be narrowed, thereby the computational cost of the model could be reduced.[34] Furthermore, four representative organic compounds involving the structures of rings (phenol and tetrahydrofuran) and chains (cyanoacetic acid and adipic acid) were selected randomly from the collected data set as case studies for validation. These compounds were not included in both the training set and test set. For the predicted target of phenol, the most relevant chemicals with similar descriptors as phenol were derived preferentially based on the weighted Euclidean distance method.

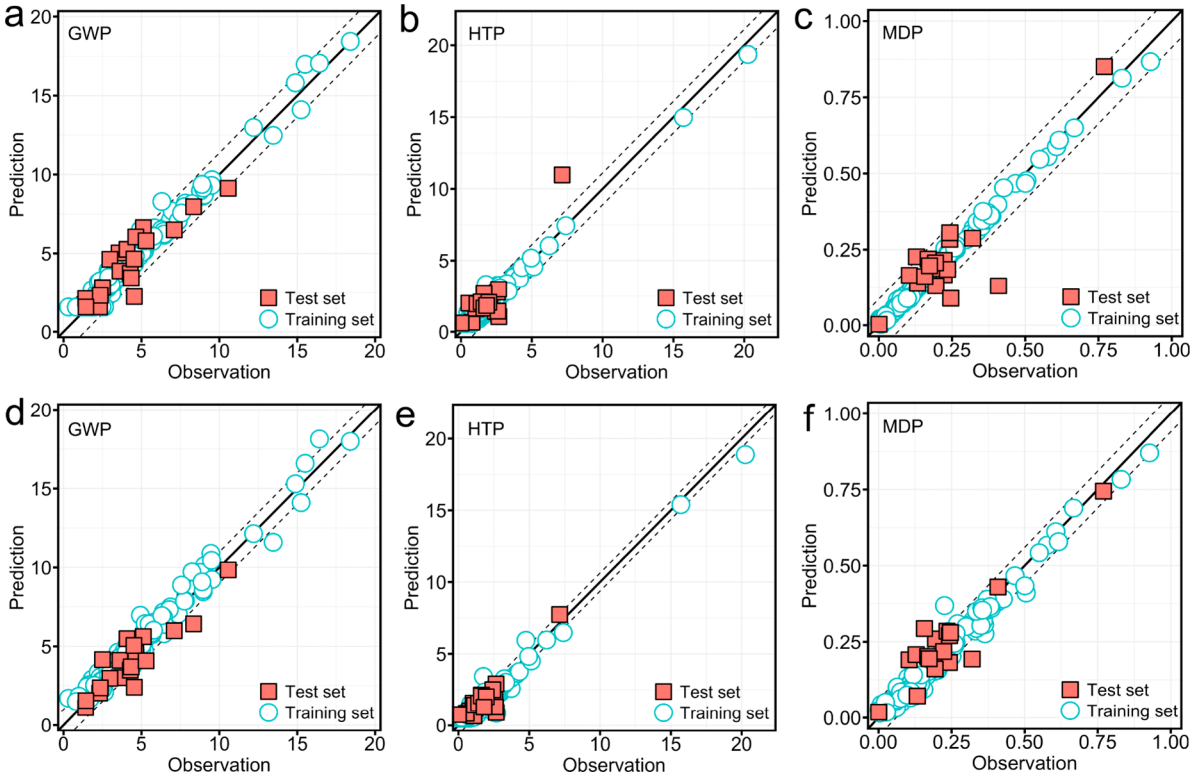


Figure 4. Prediction performance of MPN models (a–c) and improved MPEN models (d–f). The diagonal represents the perfect prediction line, and the dotted lines represent the intercepts of ± RMSE.

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444

All these screened chemicals have the structures or functional groups (*e.g.*, aromatic rings) being similar as phenol. Figure S11 illustrates the structures of the target chemical and the most relevant top ten chemicals, including aniline, sodium phenolate, cyclohexanone, toluene, methylcyclohexane, cyclohexanol, morpholine, benzaldehyde, dioxane, and alpha-picoline. Then, the most relevant 60% data selected from the original training set served as a new training set for training models to predict the life-cycle environmental impacts of phenol. Likewise, for the predicted target of tetrahydrofuran, cyanoacetic acid, and adipic acid, the most relevant training sets were also selected for MPEN model training and the top ten relevant chemicals were shown in Figures S12−S14. As shown in Table 2, the MPEN models generally yielded better predictions than those from MPN models, and the resulting GWP of phenol predicted by the MPEN model was 3.31 kg CO_2-Eq, corresponding to an absolute value of relative error of 3.1%, which is closer to the observation value of 3.21 kg CO_2-eq compared with the prediction value of the MPN model (3.43 kg CO_2-Eq). The GWP values obtained for tetrahydrofuran (6.88 kg CO_2-Eq), cyanoacetic acid (5.23 kg CO_2-Eq), and adipic acid (14.69 kg CO_2-Eq) predicted by the MPEN model were better predicted than that predicted by the MPN model (6.61 kg CO_2-Eq, 5.67 kg CO_2-eq, and 14.12 kg CO_2-Eq), accounting for the absolute value of relative error (1.3%, 2.1%, and 1.3%) to be also lower than that obtained for the MPN model (5.2%, 10.7%, and 5.1%). The prediction accuracy of the MPEN model for HTP was also improved compared with the MPN model, with the absolute value of the relative error for phenol prediction reduced from 33.3% to 10.4%, and for cyanoacetic acid, it reduced from 7.4% to 0.8%, the values being much lower than 30% reported by Song *et al.*[4] who developed ANN models by using PCA-based descriptors. Such high prediction accuracy of MPEN models validated the effectiveness of the improved models based on screening by weighted Euclidean distance, which suggested a promising potential of ANN models in predicting the life-cycle environmental impacts of unexplored chemicals.

Table 2. Prediction of Case Studies for Validation of MPEN Models

target chemicals	impact categories	observation	MPN model[a]	MPEN model[a]
phenol	GWP	3.21	3.43 (6.9%)	3.31 (3.1%)
	HTP	0.96	1.28 (33.3%)	1.06 (10.4%)
tetrahydrofuran	GWP	6.97	6.61 (5.2%)	6.88 (1.3%)
	HTP	2.54	2.79 (9.8%)	2.61 (2.8%)
cyanoacetic acid	GWP	5.12	5.67 (10.7%)	5.23 (2.1%)
	HTP	2.57	2.76 (7.4%)	2.59 (0.8%)
adipic acid	GWP	14.88	14.12 (5.1%)	14.69 (1.3%)
	HTP	1.56	1.71 (9.6%)	1.62 (3.8%)

[a]The numbers in the parentheses are the absolute value of relative error.

3.4. Model Interpretation. Machine learning models should seek for both predicting accuracy and mechanistic interpretation.[36,59] Therefore, it is also crucial to highlight the importance of molecular descriptors and their comprehension. The state-of-the-art SHAP method for interpretation was introduced by calculating Shapley values to quantify the contributions of input features. In brief, the Shapley theory

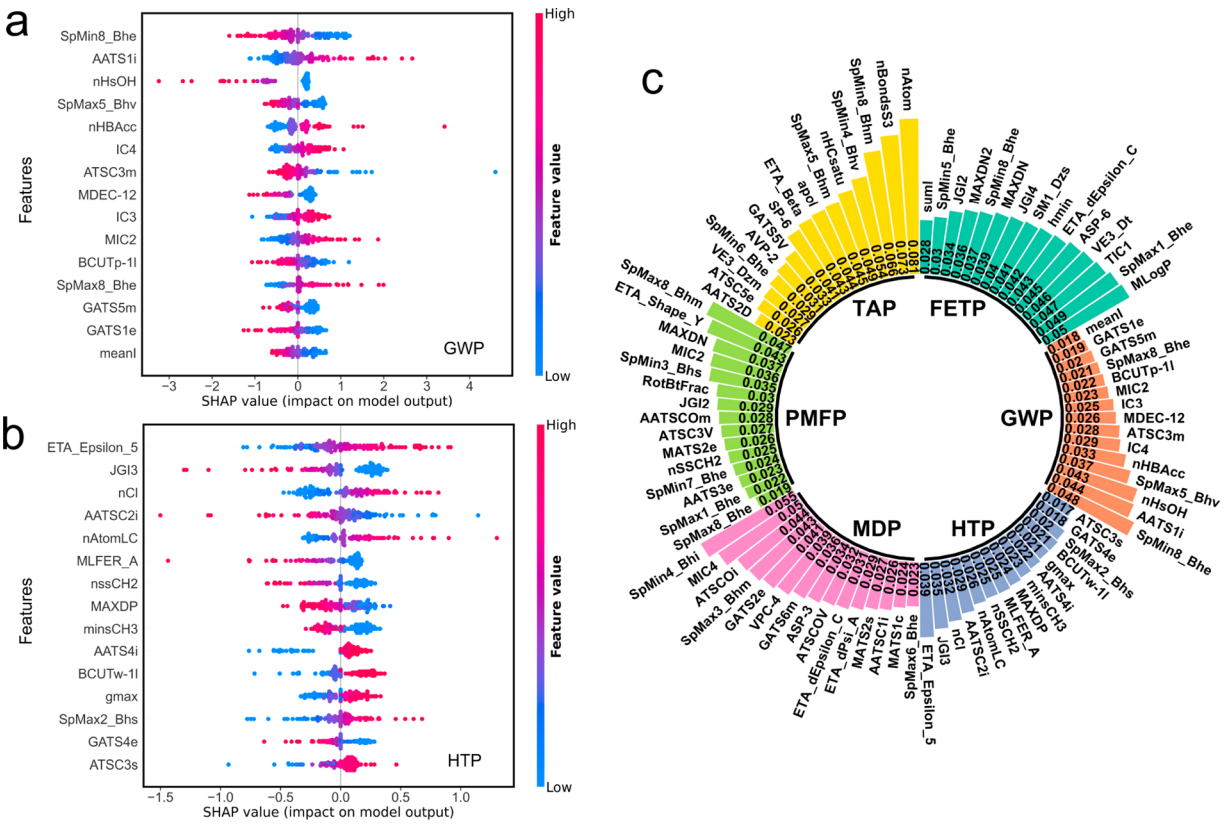


Figure 5. Shapley values for (a) GWP and (b) HTP MPEN models and (c) importance of the representative descriptors (top 15) of each impact categories.

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444

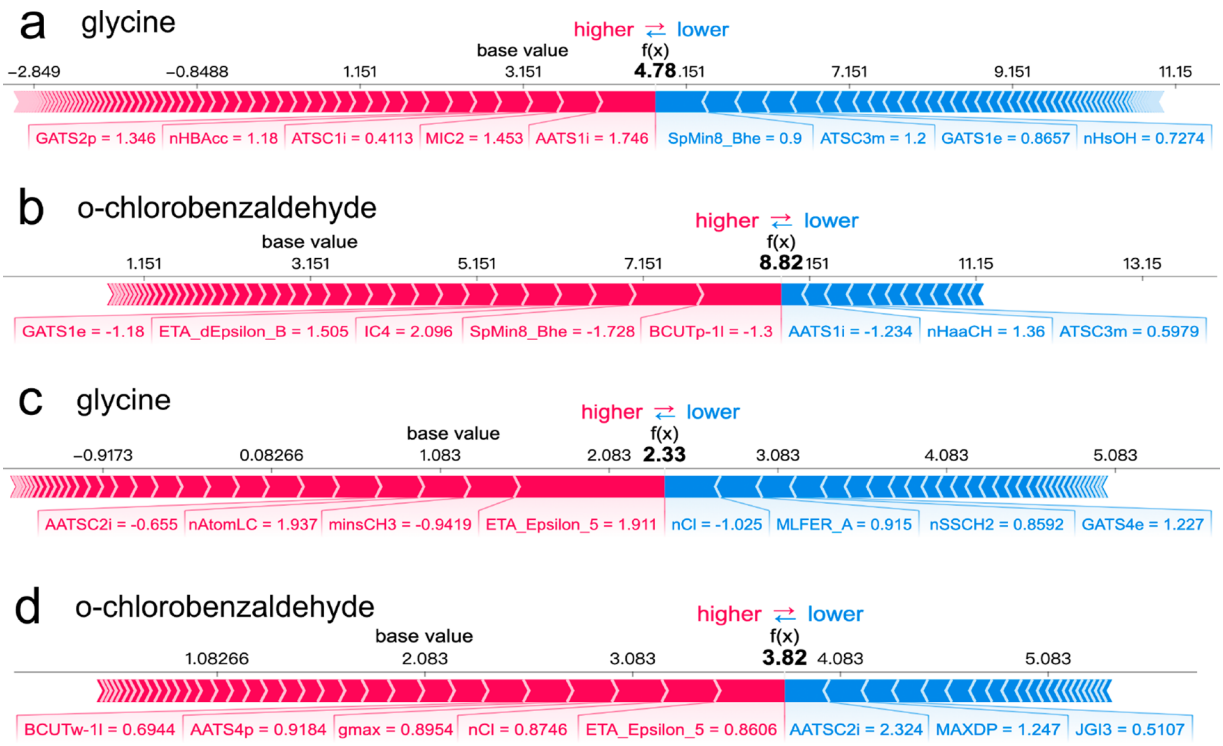


Figure 6. Shapley values of different descriptors for the GWP of glycine (a) and *o*-chlorobenzaldehyde (b) and the HTP of glycine (c) and *o*-chlorobenzaldehyde (d).

originating from the coalitional game assumes that each variable plays a role in one game and the influence of variable on final outcome is measured.[25,34,60] In this study, the SHAP theory was used with variables serving as molecular descriptors and final outcome serving as environmental impacts, respectively. Shapley values not only indicate the importance of features and identify the features that exert greater influence on the models but also demonstrate the correlation between molecular descriptors and predicted results.[59,61,62] The more positive or negative the Shapley values are, the larger positive or negative influence of the descriptor to environmental impacts, and vice versa.

To quantify contribution of the molecular descriptors in MPEN models, Shapley values of the descriptors in each model were calculated. The feature importance was measured by averaging the absolute values of each feature influence on the target variables. The detailed information on molecular descriptors was listed in the Supporting Information (Descriptors.xlsx, sheet "Descriptor description"). As shown in Figure 5a, for the GWP MPEN model, SpMin8_Bhe was the most relevant descriptor that influenced GWP, corresponding to the mean absolute Shapley value of 0.048 (Figure 5c). The high normalized feature values (red color) corresponded to negative Shapley values, and vice versa (Figure 5a), clearly indicating a strong negative correlation between SpMin8_Bhe and GWP. Likewise, AATS1i was identified to be the second relevant descriptor in the GWP MPEN model with a mean absolute Shapley value of 0.044. The high normalized feature values (red color) corresponded to positive Shapley values, where a positive correlation between AATS1i and GWP was noticed. A similar tendency was also observed for the HTP MPEN model where the ETA_Epsilon_5 was the most relevant descriptor being positively correlated to HTP (Figure 5b), corresponding to a mean absolute Shapley value of 0.039. The second relevant descriptor that influenced HTP was JGI3, accounting for a mean absolute Shapley value of 0.035, and the high normalized feature values corresponded to negative Shapley values, which demonstrated a negative correlation. Following these procedures, the importance of the top 15 representative descriptors and Shapley values for MDP, FETP, PMFP, and TAP MPEN models could also be identified (Figure S15). Accordingly, Figure 5c visualizes these descriptors with the greatest influence for each environmental impact category according to their mean absolute Shapley values.

To further verify reliability of the feature importance of the models, two chemicals (*i.e.*, glycine and *o*-chlorobenzaldehyde) were randomly selected from the Ecoinvent database. It was worth noticing that the two chemicals were not included in both the training and test sets. The Shapley values of different descriptors for the GWP and HTP of glycine and *o*-chlorobenzaldehyde were calculated. As shown in Figure 6, the length of segments in each ribbon is representative of environmental impacts contributed by descriptors. In brief, longer segment means higher contribution, and the red and blue segment denotes positive and negative contribution, respectively. The numbers in bold were the output environmental impacts predicted by MPEN models. The three descriptors that contribute most to GWP for glycine were AATS1i, SpMin8_Bhe, and MIC2, accounting for Shapley values of 1.08, −0.99, and 0.90 (Figure 6a). This result appeared to be in line with the correlation observed for the original test set (Figure 5a). As shown in Figure 6b, three descriptors that had the largest contribution to GWP of *o*-chlorobenzaldehyde were BCUTp-1l, SpMin8_Bhe, and IC4, corresponding to Shapley values of 0.92, 0.79, and 0.44, which was also consistent with the correlation shown in Figure 5a. On the basis of model interpretation, the descriptor AATS1i was identified to be positively correlated with GWP. As described in the Supporting Information (Descrip-

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444

tors.xlsx), the descriptor AATS1i was relevant to the first ionization energy, which might exert influence on the stability of the molecule and thus determine the lifetime of chemicals in the atmosphere. It was well-known that the GWP of a chemical was impacted by its capability to absorb infrared light, the wavelength range of its absorption spectrum, and the lifetime of the chemical in the atmosphere. This demonstrated the reliability of our models from a physicochemical point of view. Likewise, Figure 6c,d illustrates ETA_Epsilon_5 and nCl to be the descriptors contributing most to HTP of glycine and *o*-chlorobenzaldehyde, and the correlation was also in good agreement with the results shown in Figure 5b. Chlorine atoms involved in the structure of examined chemicals may have contribution as the dominant factor being responsible for human toxicity. Therefore, the feature importance derived from Shapley values in our models was also reliable from a physicochemical point of view, and the MPEN models were robust and meaningful for prediction and interpretation of environmental impacts. The results can be used for screening desirable chemicals or searching for chemical substitutes based on the contribution of molecular descriptors and their relevance to environmental impacts.

3.5. Implications. In this study, we have demonstrated that data processing coupled with ANN can improve the accuracy and interpretability of the models for predicting the life-cycle environmental impacts of chemicals. The data processing methods of feature selection by MI-PI and source data selection based on the weighted Euclidean distance are shown to be effective for alleviating high computation cost, increasing model prediction accuracy, and preserving the data interpretability. Compared with previously reported feature extraction method PCA,[1,4] the MI-PI method used in this study for feature selection is more favorable for model interpretation by preserving the physicochemical meanings of the original molecular descriptors without generating new variables. Meanwhile, MI-PI is useful for dealing with the data set characterized by large-volume features with a small number of available samples. We show that the model prediction accuracy can be improved through source data selection based on the weighted Euclidean distance method by screening out the chemicals most relevant to the target predicted chemicals. This will mitigate the inefficient computational cost, which is especially beneficial for the case of limited provision of available data. Our models have better performance than those reported in previous studies, especially for the GWP prediction model, indicated by R^2 of 0.81 compared with 0.48 reported by Song *et al.*[4] Moreover, interpretation of the models based on the Shapley value further demonstrates the most relevant molecular descriptors to each sort of environmental impact categories. Our models provide an initial-stage screening tool for estimating the life-cycle environmental impacts of chemicals for certain impact categories in the absence of redundant information. It will be useful for screening desirable chemicals and searching for greener substitutes in industrial processes when taking in account decreasing the environmental burdens.

Despite accuracy, robustness, and reliability of our models, there is space for further improvement for practical LCA implementations. First, since the data sets in Ecoinvent are mostly available for industrially produced chemicals, the predicted results of our models reflect the life-cycle environmental impacts of chemicals on industrial scale. Despite the models created upon the LCIA method of ReCiPe Midpoint, the framework of model building can also be extended for other LCIA methods like IPCC and CML not tested here because they also have similar characteristics of data distribution. Second, because of the limit of data availability, the number of samples for training and testing the models is relatively small, which may lead to the problem of overfitting.[56] Such an issue is expected to be addressed by an increase in the size of the commercial database as the models continue to be updated and improved in the future as more data become available. Third, the models are interpreted by posthoc analysis where different feature combinations are employed to validate performances of models.[59,61] A challenge remains with respect to interpretation of the way the molecular descriptors affect the output values, since their physicochemical meanings are unlikely to be readily interpretable. Future effort is motivated by the needed to search for descriptors that are better understood and interpreted. On the other hand, new challenge may emerge as the size of the database grows because it appears impractical to retrain the models with each new data set. In this case, it will be necessary to develop an active learning framework, which allows continuous learning along with future access to large-sized new data. Last, despite high prediction accuracy of our models obtained in this study, uncertainty of prediction still exists, which needs to be assessed in future work to further improve the model prediction accuracy and interpretability.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.2c04945.

Detailed description of the machine learning algorithms and data dimensionality reduction methods, grid search and model improvement method; detailed data of the collected chemicals; prediction performance and hyperparameters for each model; structures of the target chemical and the most relevant ten chemicals to target chemical; and other related figures for the model interpretation (PDF)

Numerical values and explanation of the molecular descriptors (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Shijie You − *State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin 150090, P. R. China;* orcid.org/0000-0001-8178-9418; Phone: +86-451-86282008; Email: sjyou@hit.edu.cn; Fax: +86-451-8628 2110

Authors

Ye Sun − *State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin 150090, P. R. China*

Xiuheng Wang − *State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin 150090, P. R. China*

Nanqi Ren − *State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin 150090, P. R. China*

Yanbiao Liu − *College of Environmental Science and Engineering, Textile Pollution Controlling Engineering Center of the Ministry of Ecology and Environment, Donghua*

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444

University, Shanghai 201620, China; orcid.org/0000-0001-8404-3806

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.2c04945

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (U21A20161, 51961125104), the Key Research and Development Program of Heilongjiang Province (Grant No. GA22B003), the Fundamental Research Funds for the Central Universities (Grant No. HIT.OCEF.2022019) and Heilongjiang Touyan Innovation Team Program (HIT-SE-01).

■ REFERENCES

(1) Zhu, X.; Ho, C.; Wang, X. Application of life cycle assessment and machine learning for high-throughput screening of green chemical substitutes. *ACS Sustainable Chem. Eng.* **2020**, *8*, 11141−11151.
(2) Poliakoff, M.; Licence, P. Sustainable technology: green chemistry. *Nature* **2007**, *450*, 810−812.
(3) Wernet, G.; Papadokonstantakis, S.; Hellweg, S.; Hungerbuhler, K. Bridging data gaps in environmental assessments: modeling impacts of fine and basic chemical production. *Green Chem.* **2009**, *11*, 1826−1831.
(4) Song, R.; Keller, A. A.; Suh, S. Rapid life-cycle impact screening using artificial neural networks. *Environ. Sci. Technol.* **2017**, *51*, 10777−10785.
(5) Arvidsson, R.; Tillman, A.-M.; Sanden, B. A.; Janssen, M.; Nordelof, A.; Kushnir, D.; Molander, S. Environmental assessment of emerging technologies: recommendations for prospective LCA. *J. Ind. Ecol.* **2018**, *22*, 1286−1294.
(6) Pesqueira, J. F. J. R.; Pereira, M. F. R.; Silva, A. Environmental impact assessment of advanced urban wastewater treatment technologies for the removal of priority substances and contaminants of emerging concern: A review. *J. Clean. Prod.* **2020**, *261*, 121078−121092.
(7) Thonemann, N.; Schulte, A.; Maga, D. How to conduct prospective life cycle assessment for emerging technologies? A systematic review and methodological guidance. *Sustainability* **2020**, *12*, 1192−1215.
(8) Pallas, G.; Vijver, M. G.; Peijnenburg, W. J. G. M.; Guinée, J. Life cycle assessment of emerging technologies at the lab scale: the case of nanowire-based solar cells. *J. Ind. Ecol.* **2020**, *24*, 193−204.
(9) Zhao, B.; Shuai, C.; Hou, P.; Qu, S.; Xu, M. Estimation of unit process data for life cycle assessment using a decision tree-based approach. *Environ. Sci. Technol.* **2021**, *55*, 8439−8446.
(10) Hou, P.; Cai, J.; Qu, S.; Xu, M. Estimating missing unit process data in life cycle assessment using a similarity-based approach. *Environ. Sci. Technol.* **2018**, *52*, 5259−5267.
(11) Liao, M.; Kelley, S.; Yao, Y. Generating energy and greenhouse gas inventory data of activated carbon production using machine learning and kinetic based process simulation. *ACS Sustainable Chem. Eng.* **2020**, *8*, 1252−1261.
(12) Parvatker, A. G.; Eckelman, M. J. Comparative evaluation of chemical life cycle inventory generation methods and implications for life cycle assessment results. *ACS Sustainable Chem. Eng.* **2019**, *7*, 350−367.
(13) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokonstantakis, S.; Hungerbuhler, K. Molecular-structure-based models of chemical inventories using neural networks. *Environ. Sci. Technol.* **2008**, *42*, 6717−6722.
(14) Kaab, A.; Sharifi, M.; Mobli, H.; Nabavi-Pelesaraei, A.; Chau, K. W. Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. *Sci. Total Environ.* **2019**, *664*, 1005−1019.
(15) Nabavi-Pelesaraei, A.; Rafiee, S.; Mohtasebi, S. S.; Hosseinzadeh-Bandbafha, H.; Chau, K. W. Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production. *Sci. Total Environ.* **2018**, *631*, 1279−1294.
(16) Tsoy, N.; Steubing, B.; van der Giesen, C.; Guinee, J. Upscaling methods used in ex-ante life cycle assessment of emerging technologies: a review. *Int. J. Life Cycle Ass.* **2020**, *25*, 1680−1692.
(17) Calvo-Serrano, R.; González-Miquel, M.; Papadokonstantakis, S.; Guillén-Gosálbez, G. Predicting the cradle-to-gate environmental impact of chemicals from molecular descriptors and thermodynamic properties via mixed-integer programming. *Comput. Chem. Eng.* **2018**, *108*, 179−193.
(18) Eckelman, M. J. Life cycle inherent toxicity: a novel LCA-based algorithm for evaluating chemical synthesis pathways. *Green Chem.* **2016**, *18*, 3257−3264.
(19) Hou, P.; Jolliet, O.; Zhu, J.; Xu, M. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environ. Int.* **2020**, *135*, 105393−105405.
(20) Marvuglia, A.; Kanevski, M.; Benetto, E. Machine learning for toxicity characterization of organic chemical emissions using USEtox database: learning the structure of the input space. *Environ. Int.* **2015**, *83*, 72−85.
(21) Raies, A. B.; Bajic, V. B. *In silico* toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6*, 147−172.
(22) Ghamri, M.; Harkati, D.; Belaidi, S.; Boudergua, S.; Said, R. B.; Linguerri, R.; Chambaud, G.; Hochlaf, M. Carbazole derivatives containing chalcone analogues targeting topoisomerase II inhibition: first principles characterization and QSAR modelling. *Spectrochim. Acta, Part A* **2020**, *242*, 118724−118733.
(23) Khanali, M.; Mobli, H.; Hosseinzadeh-Bandbafha, H. Modeling of yield and environmental impact categories in tea processing units based on artificial neural networks. *Environ. Sci. Pollut. R.* **2017**, *24*, 26324−26340.
(24) Ozbilen, A.; Aydin, M.; Dincer, I.; Rosen, M. A. Life cycle assessment of nuclear-based hydrogen production via a copper-chlorine cycle: a neural network approach. *Int. J. Hydrogen Energy* **2013**, *38*, 6314−6322.
(25) 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*, 1854−1863.
(26) Khoshnevisan, B.; Rafiee, S.; Omid, M.; Mousazadeh, H.; Sefeedpari, P. Prognostication of environmental indices in potato production using artificial neural networks. *J. Clean. Prod.* **2013**, *52*, 402−409.
(27) Sharif, S. A.; Hammad, A. Developing surrogate ANN for selecting near-optimal building energy renovation methods considering energy consumption, LCC and LCA. *J. Build. Eng.* **2019**, *25*, 100790−100805.
(28) Wu, Y.; Wang, G. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. *Int. J. Mol. Sci.* **2018**, *19*, 2358−2377.
(29) Servien, R.; Latrille, E.; Patureau, D.; Hélias, A. Machine learning models based on molecular descriptors to predict human and environmental tox icological factors in continental freshwater. *Peer Community Journal* **2022**, *2*, No. e15.
(30) Lysenko, A.; Sharma, A.; Boroevich, K. A.; Tsunoda, T. An integrative machine learning approach for prediction of toxicity-related drug safety. *Life Sci. Alliance* **2018**, *1*, No. e201800098.
(31) Marvuglia, A.; Leuenberger, M.; Kanevski, M.; Benetto, E. Random Forest for toxicity of chemical emissions: features selection and uncertainty quantification. *J. Environ. Account. Ma.* **2015**, *3*, 229−241.
(32) Marvuglia, A.; Kanevski, M.; Leuenberger, M.; Benetto, E. Variables selection for ecotoxicity and human toxicity characterization using Gamma Test. *Computational Science and Its Applications* **2014**, *8581*, 640−652.

(33) Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16−28.
(34) Zhang, K.; Zhong, S.; Zhang, H. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. *Environ. Sci. Technol.* **2020**, *54*, 7008−7018.
(35) Effrosynidis, D.; Arampatzis, A. An evaluation of feature selection methods for environmental data. *Ecol. Inform.* **2021**, *61*, 101224−101234.
(36) Altmann, A.; Tolosi, L.; Sander, O.; Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340−1347.
(37) Zhu, Z.; Ong, Y. S.; Dash, M. Wrapper-filter feature selection algorithm using a memetic framework. *IEEE Trans. Syst. Man. Cy. B* **2007**, *37*, 70−76.
(38) Hu, Z.; Bao, Y.; Xiong, T.; Chiong, R. Hybrid filter-wrapper feature selection for short-term load forecasting. *Eng. Appl. Artif. Int.* **2015**, *40*, 17−27.
(39) Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: a new perspective. *Neurocomputing* **2018**, *300*, 70−79.
(40) Sebban, M.; Nock, R. A hybrid filter/wrapper approach of feature selection using information theory. *Pattern Recogn.* **2002**, *35*, 835−846.
(41) Wu, X.; Zhu, X.; Wu, G.; Ding, W. Data mining with big data. *IEEE T. Knowl. Data En.* **2014**, *26*, 97−108.
(42) Wang, B.; Liu, X.; Yu, B.; Jia, R.; Gan, X. An improved WiFi positioning method based on fingerprint clustering and signal weighted Euclidean distance. *Sensors* **2019**, *19*, 2300−2320.
(43) Greenacre, M. Ordination with any dissimilarity measure: a weighted Euclidean solution. *Ecology* **2017**, *98*, 2293−2300.
(44) Doost, R.; Sayadian, A.; Shamsi, H. A new perceptually weighted distance measure for vector quantization of the STFT amplitudes in the speech application. *IEICE Electron. Expr.* **2009**, *6*, 824−830.
(45) Wernet, G.; Bauer, C.; Steubing, B.; Reinhard, J.; Moreno-Ruiz, E.; Weidema, B. The ecoinvent database version 3 (part I): overview and methodology. *Int. J. Life Cycle Ass.* **2016**, *21*, 1218−1230.
(46) Steubing, B.; Wernet, G.; Reinhard, J.; Bauer, C.; Moreno-Ruiz, E. The ecoinvent database version 3 (part II): analyzing LCA results and comparison to version 2. *Int. J. Life Cycle Ass.* **2016**, *21*, 1269−1281.
(47) Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.
(48) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(49) Wu, D.; Zhang, D.; Liu, S.; Jin, Z.; Chowwanonthapunya, T.; Gao, J.; Li, X. Prediction of polycarbonate degradation in natural atmospheric environment of China based on BP-ANN model with screened environmental factors. *Chem. Eng. J.* **2020**, *399*, 125878−125889.
(50) Xie, Y.; Zhang, C.; Hu, X.; Zhang, C.; Kelley, S. P.; Atwood, J. L.; Lin, J. Machine learning assisted synthesis of metal-organic nanocapsules. *J. Am. Chem. Soc.* **2020**, *142*, 1475−1481.
(51) Ciaburro, G.; Iannace, G.; Passaro, J.; Bifulco, A.; Marano, A. D.; Guida, M.; Marulo, F.; Branda, F. Artificial neural network-based models for predicting the sound absorption coefficient of electrospun poly(vinyl pyrrolidone)/silica composite. *Appl. Acoust.* **2020**, *169*, 107472−107479.
(52) Zhu, T.; Chen, W.; Gu, Y.; Jafvert, C. T.; Fu, D. Polyethylene-water partition coefficients for polychlorinated biphenyls: Application of QSPR predictions models with experimental validation. *Water Res.* **2021**, *207*, 117799−117807.
(53) Mirsoleimaniazizi, S. M.; Amooey, A. A.; Ghasemi, S.; Salkhordehpanbechouleh, S. Modeling the removal of endosulfan from aqueous solution by electrocoagulation process using artificial neural network (ANN). *Ind. Eng. Chem. Res.* **2015**, *54*, 9844−9849.
(54) Aquilina, N. J.; Delgado-Saborit, J. M.; Bugelli, S.; Ginies, J. P.; Harrison, R. M. Comparison of machine learning approaches with a general linear model to predict personal exposure to benzene. *Environ. Sci. Technol.* **2018**, *52*, 11215−11222.
(55) Defernez, M.; Kemsley, E. K. Avoiding overfitting in the analysis of high-dimensional data in with artificial neural networks (ANNs). *Analyst* **1999**, *124*, 1675−1681.
(56) Bejani, M. M.; Ghatee, M. A systematic review on overfitting control in shallow and deep neural networks. *Artif. Intell. Rev.* **2021**, *54*, 6391−6438.
(57) Huang, N.; Lu, G.; Xu, D. A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. *Energies* **2016**, *9*, 767−791.
(58) Nembrini, S. On the behaviour of permutation-based variable importance measures in random forest clustering. *J. Chemom.* **2019**, *33*, No. e3135.
(59) Zhang, Y.; Xie, C.; Xue, L.; Tao, Y.; Yue, G.; Jiang, B. A post-hoc interpretable ensemble model to feature effect analysis in warfarin dose prediction for Chinese patients. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 840−851.
(60) Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. ″pySiRC″: machine learning combined with molecular fingerprints to predict the reaction rate constant of the radical-based oxidation processes of aqueous organic contaminants. *Environ. Sci. Technol.* **2021**, *55*, 12437−12448.
(61) Li, L.; Qiao, J.; Yu, G.; Wang, L.; Li, H. Y.; Liao, C.; Zhu, Z. Interpretable tree-based ensemble model for predicting beach water quality. *Water Res.* **2022**, *211*, 118078−118090.
(62) Maulana Kusdhany, M. I.; Lyth, S. M. New insights into hydrogen uptake on porous carbon materials via explainable machine learning. *Carbon* **2021**, *179*, 190−201.

https://doi.org/10.1021/acs.est.2c04945
Environ. Sci. Technol. 2023, 57, 3434−3444