Supporting Information

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

Ye Sun [a], Xiuheng Wang [a], Nanqi Ren [a], Yanbiao Liu [b], Shijie You [a,] *

[a] *State Key Laboratory of Urban Water Resource and Environment, School of Environment, Harbin Institute of Technology, Harbin 150090, P. R. China*

[b] *College of Environmental Science and Engineering, Textile Pollution Controlling Engineering Center of the Ministry of Ecology and Environment, Donghua University, Shanghai 201620, China.*

Corresponding author:
* Shijie You
P. O. Box 2603#, No. 73, Huanghe Road, Nangang District, Harbin, 150090, China.
Tel.: +86-451-86282008; Fax: +86-451-8628 2110; orcid.org/0000-0001-8178-9418
E-mail: sjyou@hit.edu.cn

The supporting information (a total of 33 pages) contains:

Texts S1-S4

Figures S1-S15

Tables S1-S6

Text S1. Briefly description of the machine learning algorithms

1) Random Forest (RF)

As a expand of bagging, random forest is a method that contains multiple decision trees, and its output is determined by the mode of individual tree output.[1] On the basis of building Bagging ensemble with decision tree as the base learner, the samples of the decision tree generated by the RF are randomly generated and the feature values of the decision tree are randomly selected. The convergence of random forest is similar to bagging.[2, 3] When only one base learner is included, the performance of RF is often poor. With the increase of the number of learners, the performance of random forest will be significantly enhanced.

RF model is trained using scikit-learn in Python 3.7 and optimized with different hyper-parameters including numbers of base estimators, depth of the tree, the number of samples required to split an internal node and the number of samples required to be at a leaf node.

2) Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized distributed gradient enhancement library designed to be efficient, flexible, and portable.[4] It implements the machine learning algorithm in the framework of gradient propulsion. XGBoost provides a parallel gradient boosting decision tree (also known as GBDT), which can quickly and accurately solve many data science problems. Its basic idea is to form a strong learner through the combination of weak learners. It solves the weighted quantiles by training a series of weak learners, fitting these weak learners to the training data and making predictions by computing the weighted average.[5] Specifically, XGBoost improves the algorithm based on gradient boosting decision tree and adds regularization to the objective function to control the complexity of the model and prevent over fitting.

In this study, XGBoost model is trained using XGboost library in Python 3.7 and optimized with different hyper-parameters including learning rate, numbers of base trees, depth of the tree and sampling rate of samples.

3) Support Vector Machine (SVM)

SVM is a data mining method based on statistical learning theory.[1] It can deal with many problems, such as regression problem, classification problem and discriminant analysis. The pivotal role of SVM method is to solve nonlinear problems. SVM maps the sample space to a high-dimensional feature space with a nonlinear kernel function. In the case of regression, it is necessary to introduce an insensitive loss function, which means that if the absolute value of the difference between the predicted value and the actual value is less than the threshold, the loss does not need to be calculated, although the predicted value and the observed value may not be exactly equal.

SVM model was trained using scikit-learn in Python 3.7 and investigated different kernel functions.

4) Artificial Neural Network (ANN)

Artificial neural networks simulate the way biological nervous systems process information.[6] It reflects many basic characteristics of human brain function and is a highly complex nonlinear dynamic learning system.[7] It is especially suitable for dealing with imprecise and fuzzy information processing problems that need to consider many factors and conditions at the same time. Each layer of neurons and the next layer of neurons are fully interconnected, and there is neither the same layer connection nor cross layer connection between neurons.[8] Such a neural network structure is called multilayer feedforward neural network. The feedforward neural network consists of an input layer, multiple hidden layers and an output layer, and each layer contains a different number of neurons.[9, 10] The input layer neurons receive external input, the hidden layer and output layer neurons process the data through the activation function, and the final result is output by the output layer neurons. The performance of neural network model is directly related to the optimal structure and hyper parameters.

An ANN model is trained using scikit-learn in Python 3.7 and usually consists of three parts: an input layer, some hidden layers and an output layer. The key hyper-parameters including the number of hidden layers, the number of neurons in hidden layers, the activation functions and the optimizers were optimized by using the testing data sets to ensure that the best ANN model was obtained.

Text S2. Description of the data dimensionality reduction methods

1) Principal component analysis (PCA).

PCA is a statistical multivariant method for data dimensionality reduction, which uses orthogonal transformation to convert a series of possible linearly related variables into a group of linearly uncorrelated new variables, also known as principal components (PC), so as to use the new variables to show the characteristics of data in smaller dimensions. [11] PCA method can extract information from a high-dimensional space by projecting it onto a lower-dimensional subspace and it is a useful method for developing predictive machine learning models. The main drawback of this method is these new principal components variables can hardly provide us the physical meaning of the original descriptors, and therefore it is difficult to explain the new variables. [12] In this study, we used the features extracted by PCA that preserve 95% of the variances in the preprocessed data subset.

2) Mutual information (MI).

Mutual information (MI) is originated from information theory that can measure the amount of information about one random variable X to another random variable Y. The degree of uncertainty in the variable X [$H(X)$] is related to the probability distribution of X [$P_X(x)$] expressed in Eq. 1. The joint entropy of two random variables X and Y is defined as $H(X,Y)$ expressed in Eq. 2, where the $P_{X,Y}(x,y)$ is the joint probability distribution of variables X and Y. The decrease of uncertainty after observing Y is defined as MI(X, Y).[13]

$$H(X)=-\sum_{x\in X} P_X(x)\log P_X(x) \tag{1}$$

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y} P_{X,Y}(x,y)\log P_X(x,y) \tag{2}$$

$$\mathrm{MI}(X,Y)=H(X)+H(Y)-H(X,Y) \tag{3}$$

Clearly visible is that a low MI value indicates the variable to be nearly independent. Due to a lower computational cost, MI is useful for selecting the relevant features in machine learning models. Here, we screened out the features with MI value greater than 0.01 between the features and the corresponding environmental impact results.

3) Permutation importance (PI).

PI is a heuristic approach used for calculating feature importance.[13] The importance of the features is arranged according to the impacts on model performance by shuffling them. Briefly, the model was first validated using the original training set, and then was validated again by shuffling one of the features in the training set. If the model performance after feature shuffling is lower than that before feature shuffling, the feature is of high importance, otherwise, the feature is of low importance. The PI of the features can be expressed as:

$$\mathrm{PI}(X_i)=s-\frac{1}{K}\sum_{k=1}^{K}s_{k,X_i} \tag{4}$$

where $\mathrm{PI}(X_i)$ indicates the importance of feature X_i, s represents the R^2 of the model predicted with the training set, s_{k,X_i} represents the R^2 of the model predicted with the training set after the k^{th} shuffle of feature X_i, and K is the number of times to randomly shuffle feature X. In this study, the value of K was 10. According to the $\mathrm{PI}(X_i)$ value, the most important features with $\mathrm{PI}(X_i)>0$ were screened out for model training.

4) Mutual information-permutation importance (MI-PI).

To improve model performance and interpretability, a feature selection method based on mutual information and permutation importance (MI-PI) was proposed. The procedure of this method is described as follows:

(1) MI method was used at first to filter out the redundant and irrelevant features.

(2) The features selected by the MI method in the previous step were used to build the model.

(3) The features were further selected using PI method based on the model built in Step (2).

(4) Rebuilt the model with the features obtained in Step (3).

Text S3. Grid search

Grid search is a widely used method to obtain optimal ANN structures.[14, 15] In this study, the parameters used included the learning rate (0.0001, 0.001, 0.01), the number of hidden layers (1, 2, 3), the number of neurons in the hidden layers (16, 32, 64, 128, 256), the activation function (‘LeakyReLU’, ‘tanh’, ‘softmax’), the optimizer function (‘Adam’, ‘SGD’), the parameter initial modes (‘he_uniform’, ‘he_normal’) and the batch size (32, 64, 128, 256). A total of 2160 grid points were assessed for each life cycle environmental impact category, and the best configuration of ANNs was selected for each impact category (Table S3) based on the prediction results.

Text S4. Model improvement based on Euclidean distance

During model training, similarities of the feature ranges could be observed between the training set and the test set if the models worked well. When the model prediction was satisfactory (*e. g.*

R^2=0.75 and 0.70 for GWP and HTP model), the ranges of feature values for training and test sets were similar (Figure S8a and S9a). Adversely, if the model prediction was relatively poor (R^2=0.66 and 0.62 for GWP and HTP model), the ranges of feature values for training and test sets diverged greatly (Figure S8b and S9b). Therefore, a smaller-scale training set could be used to well predict the given target chemical to be predicted if they were of similar data ranges. A detailed description was reported by Zhang et al.[14]

To quantify similarity of the datasets, an algorithm is needed to measure the distances between adjacent data points, which can be achieved by Euclidean distance.[16] On the other hand, however, given that each feature has different contribution in our models, it appears unlikely to quantify the similarity of these data using Euclidean distance directly. For example, for the given samples S_i=(feature 1, feature 2) by assuming S_1=(0.4, 0.4), S_2=(0.2, 0.5), S_3=(0.1, 0.9), if the similarity is calculated directly by Euclidean distance, the sample S_1 is more similar as S_2 than S_3. However, if feature 1 contributed 90% while feature 2 contributed only 10%, S_3 would be the sample being more similar as S_2 rather than S_1. To address this issue, weighted Euclidean distance[17] was more suitable to quantify the similarity of dataset in this study. Since the PI method can measure the feature importance under the guidance of model prediction accuracy, the weight of the feature (w_i) is defined as

$$w_i = \frac{PI(X_i)}{\sum_{i=1}^{n} PI(X_i)} \tag{5}$$

where $PI(X_i)$ is the PI value of the i^{th} feature. The weighted Euclidean distance is more advantageous because the inherent contribution of each feature to the models can be quantified by weighting, which avoided the interference of irrelevant features on evaluation of data similarity.

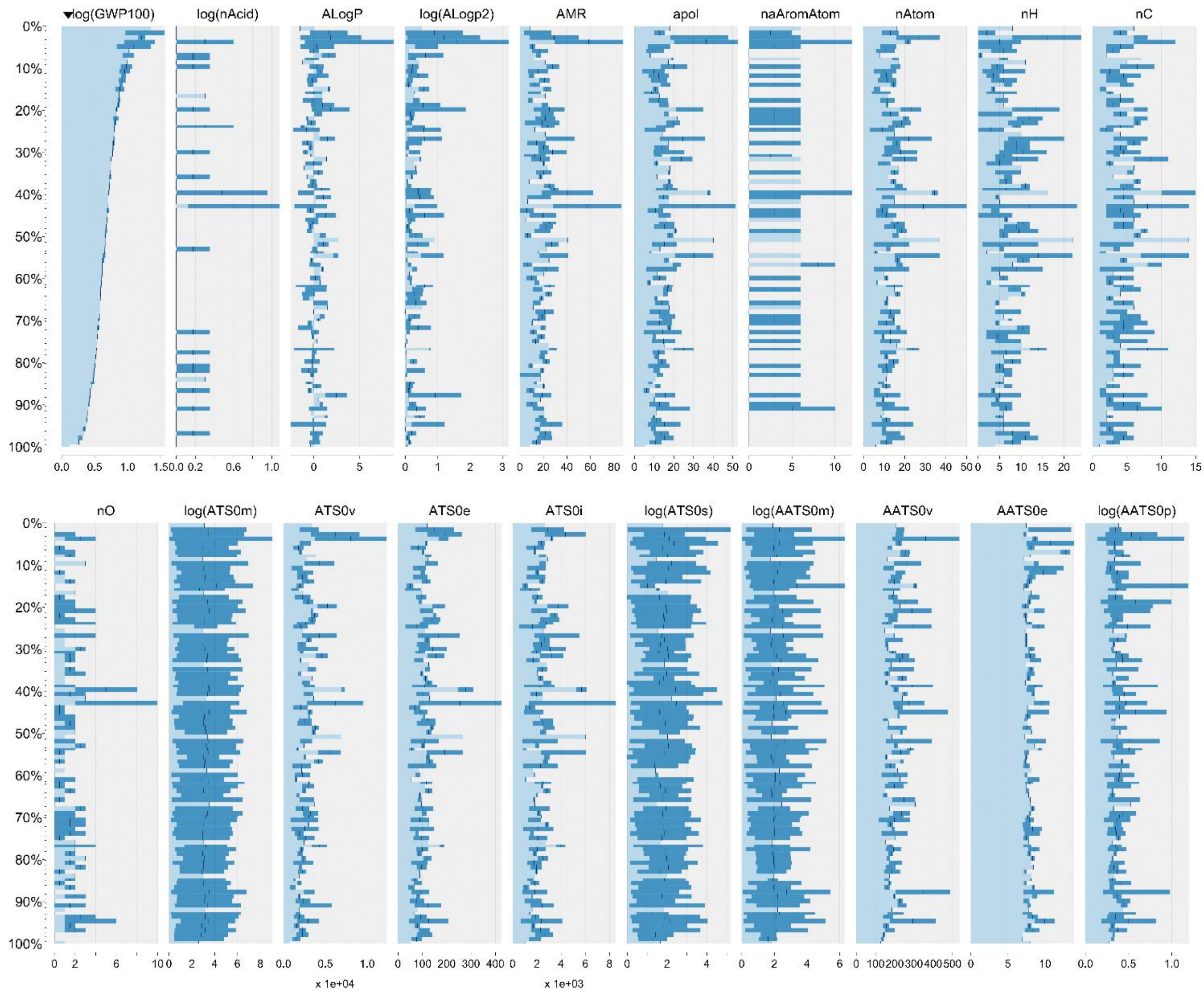

Figure S1. Visualization of data distribution of the descriptors according to the descending order of GWP.

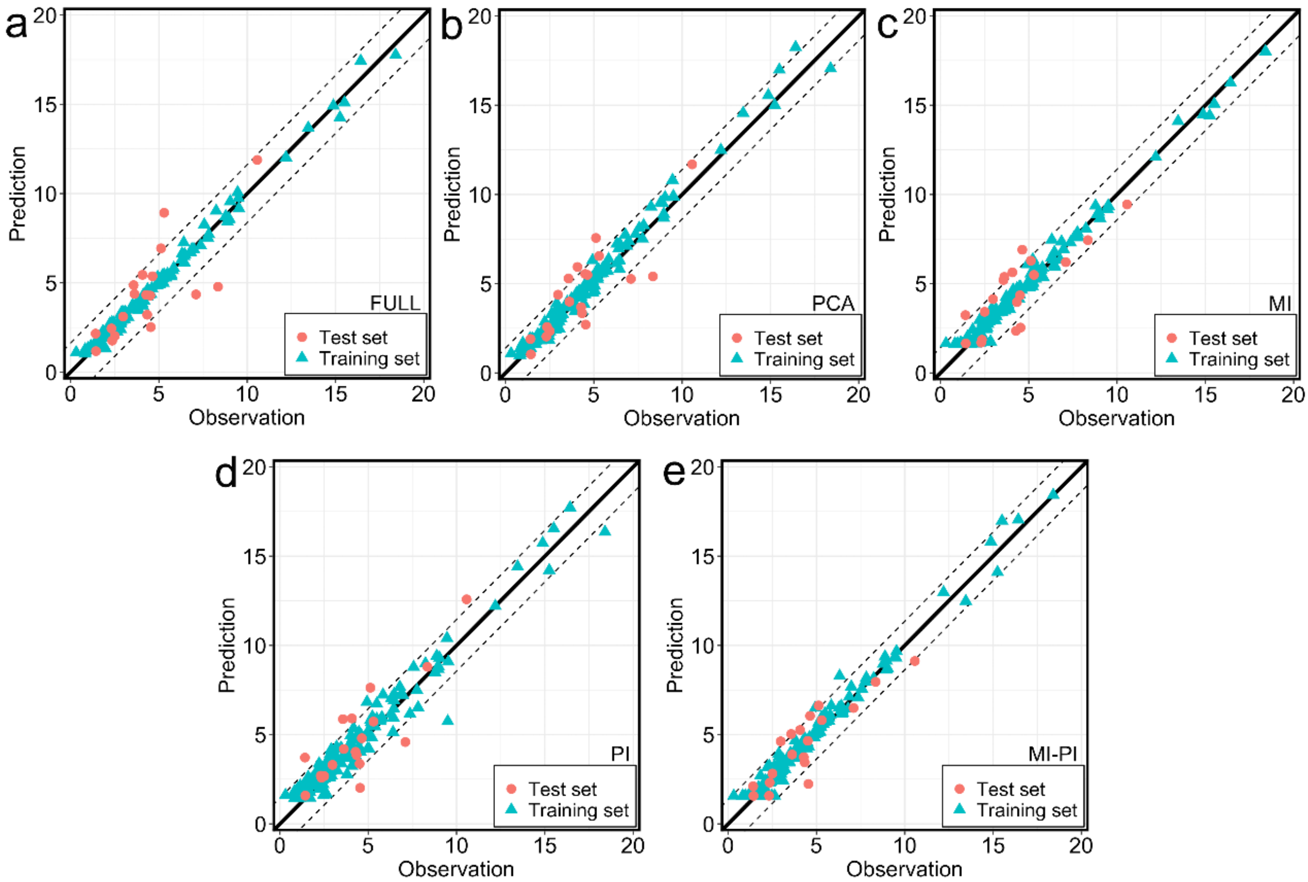

Figure S2. Prediction performance of ANN models for GWP based on five different data dimensionality reduction approaches. The diagonal represents the perfect prediction line, and the dotted lines represent the intercepts of ±RMSE.

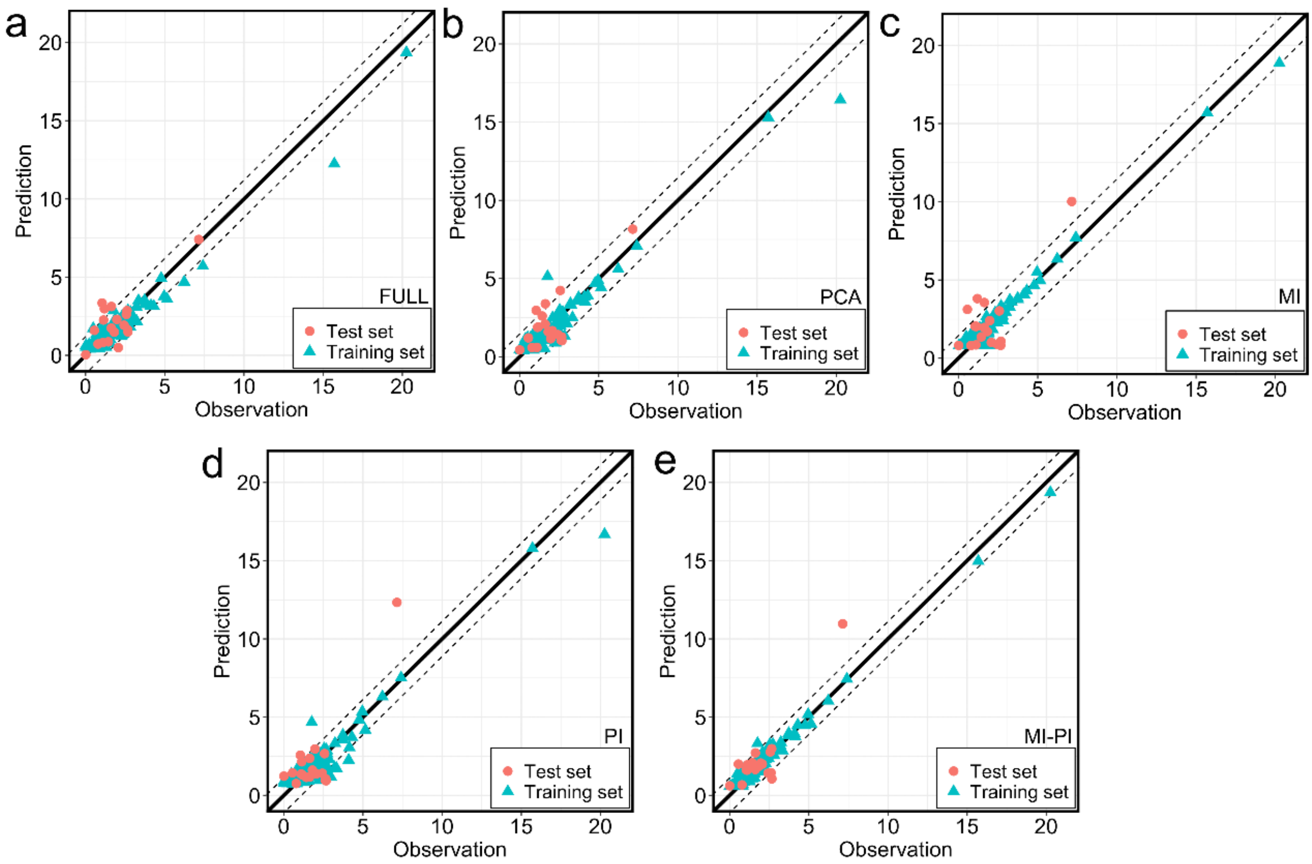

Figure S3. Prediction performance of ANN models for HTP based on five different data dimensionality reduction approaches. The diagonal represents the perfect prediction line, and the dotted lines represent the intercepts of ±RMSE.

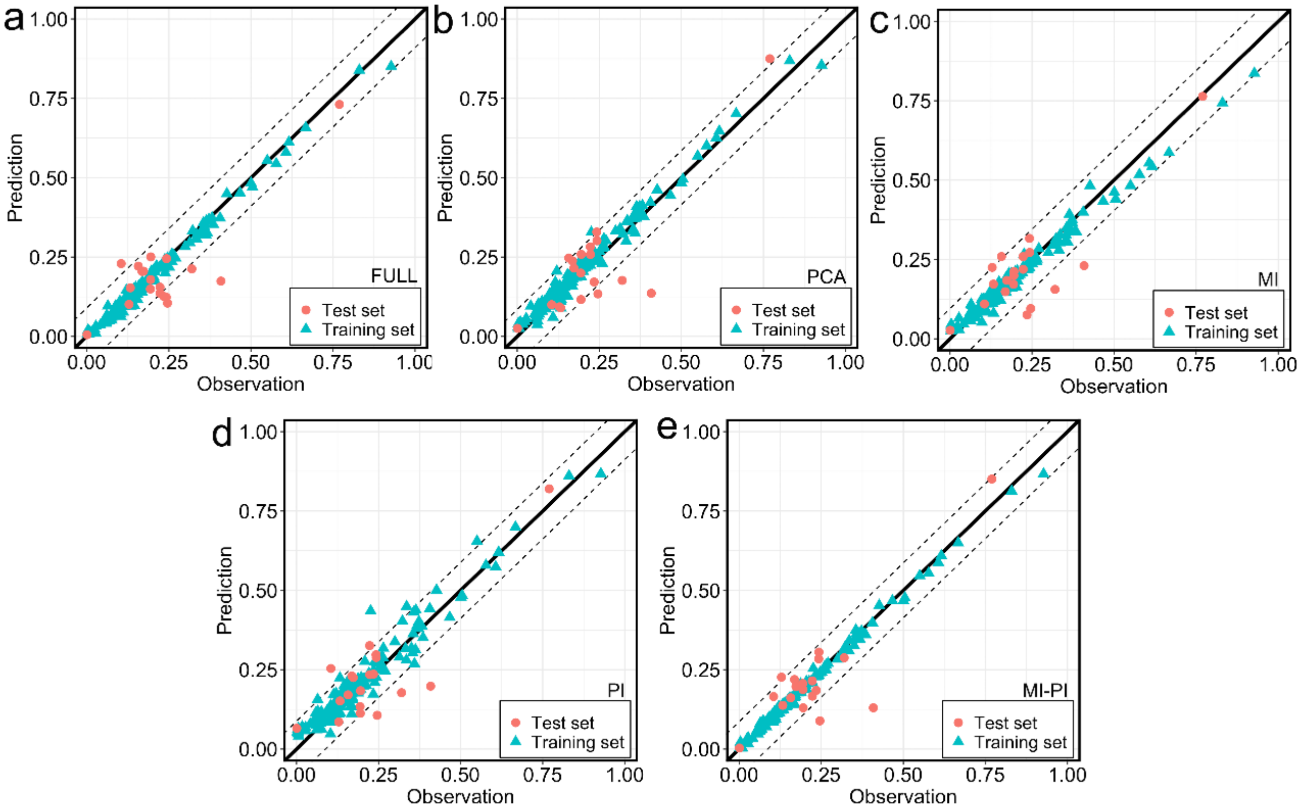

Figure S4. Prediction performance of ANN models for MDP based on five different data dimensionality reduction approaches. The diagonal represents the perfect prediction line, and the dotted lines represent the intercepts of ±RMSE.

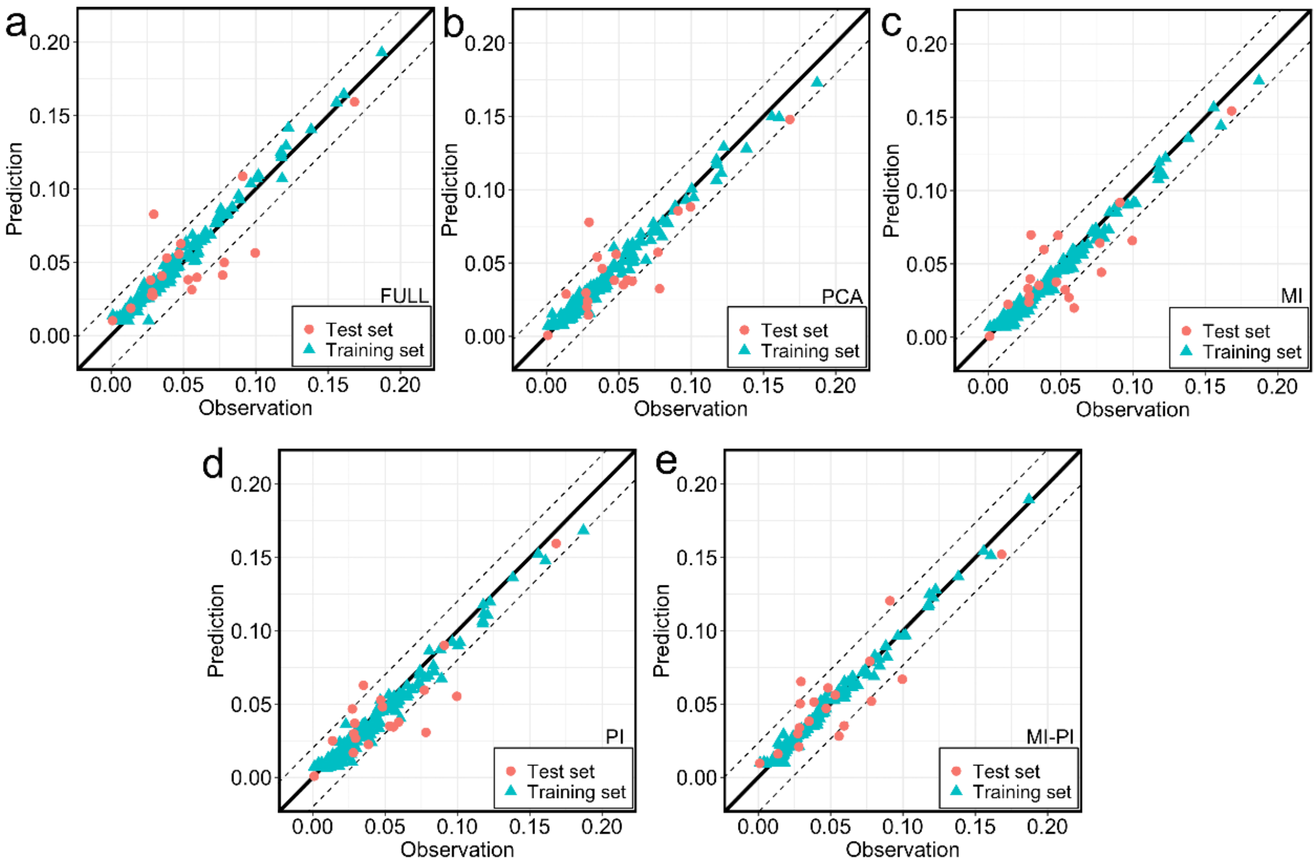

Figure S5. Prediction performance of ANN models for FETP based on five different data dimensionality reduction approaches. The diagonal represents the perfect prediction line, and the dotted lines represent the intercepts of ±RMSE.

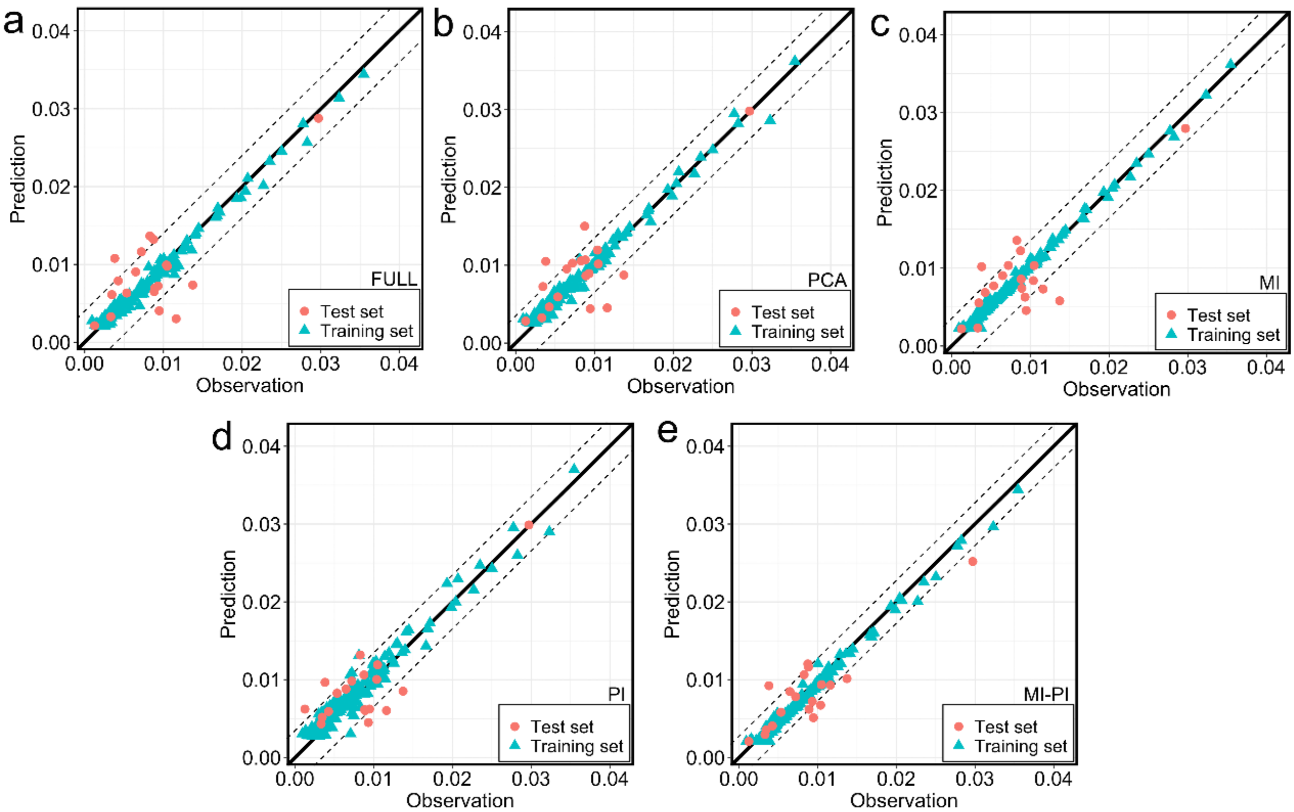

Figure S6. Prediction performance of ANN models for PMFP based on five different data dimensionality reduction approaches. The diagonal represents the perfect prediction line, and the dotted lines represent the intercepts of ±RMSE.

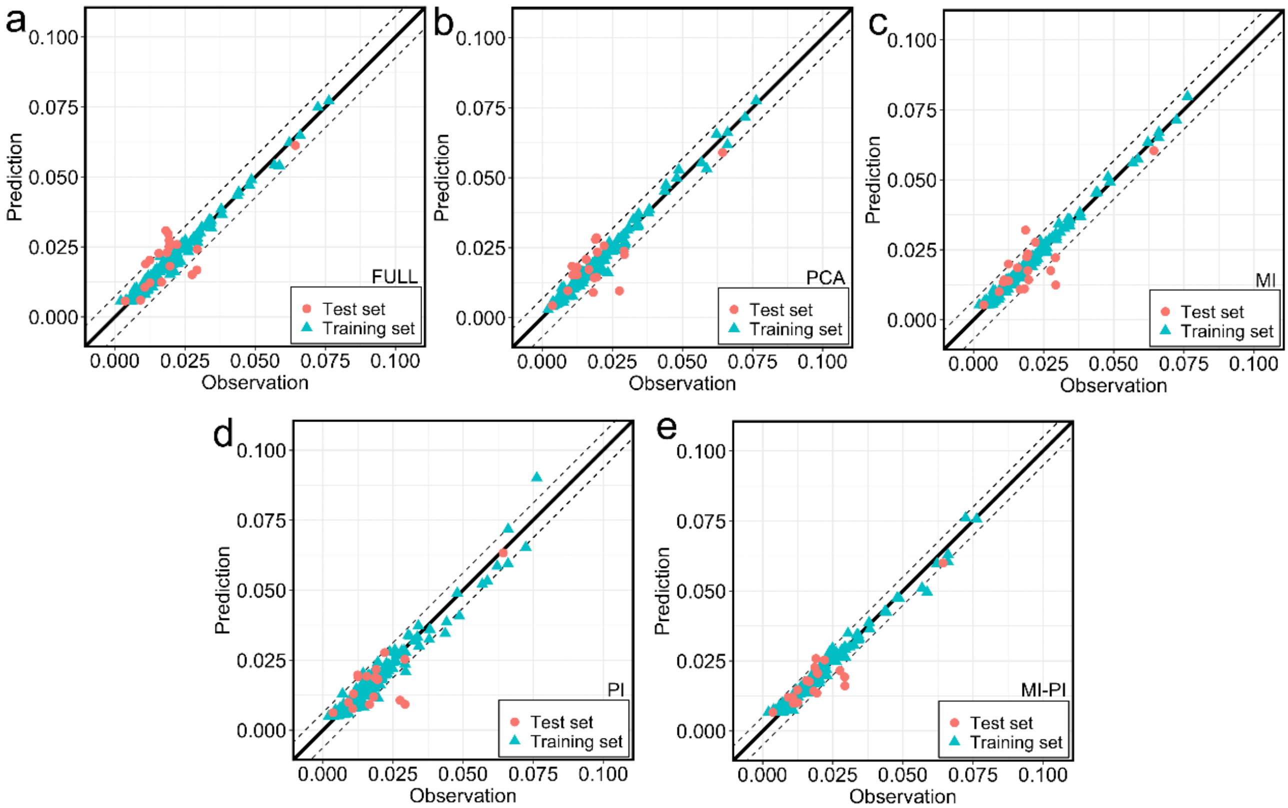

Figure S7. Prediction performance of ANN models for TAP based on five different data dimensionality reduction approaches. The diagonal represents the perfect prediction line, and the dotted lines represent the intercepts of ±RMSE.

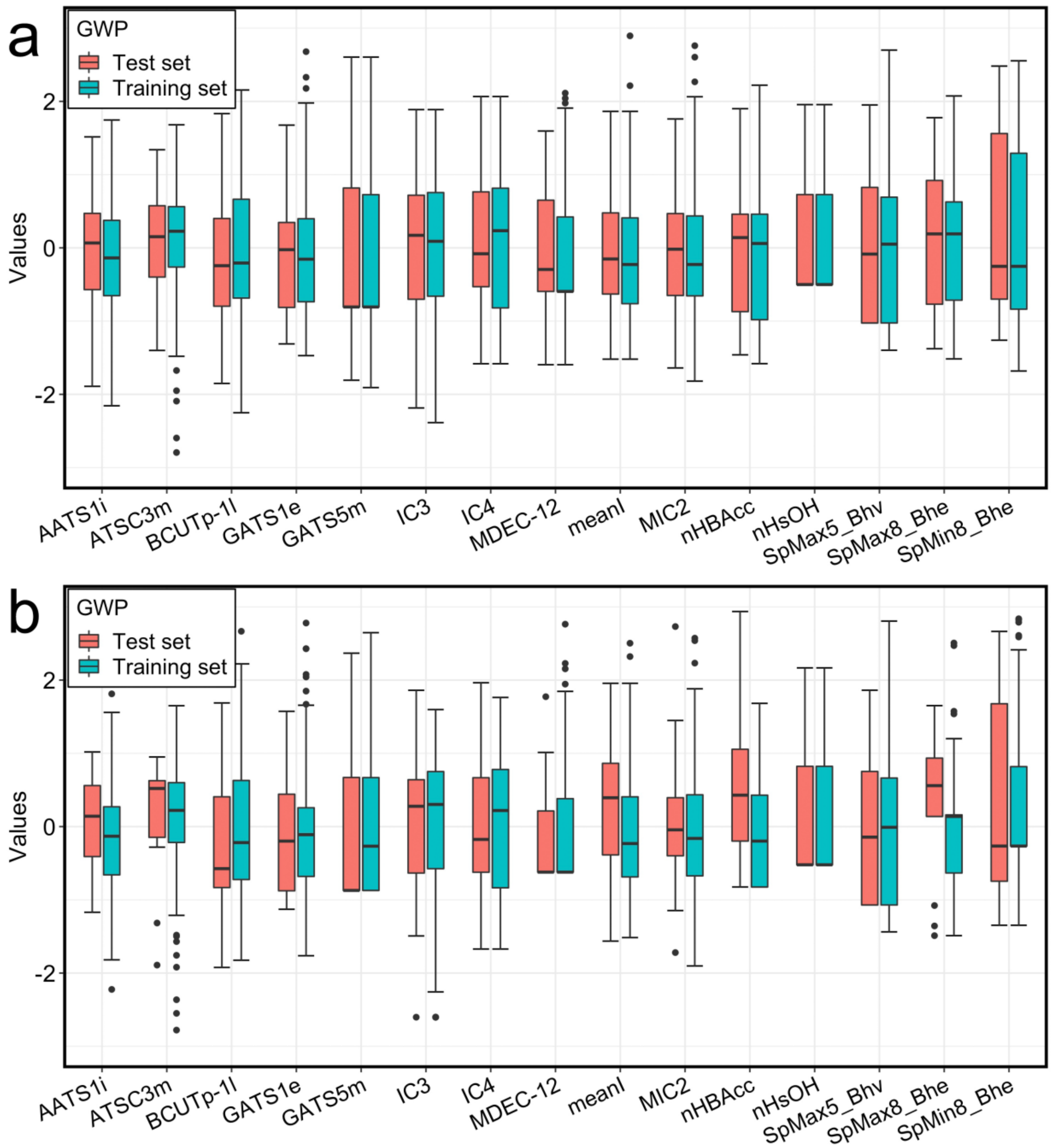

Figure S8. Ranges of the input feature values (after normalized) for training and test data points in GWP models under two scenarios: (a) a good predictive model (test set $R^2 = 0.75$) and (b) a poor model (test set $R^2 = 0.53$).

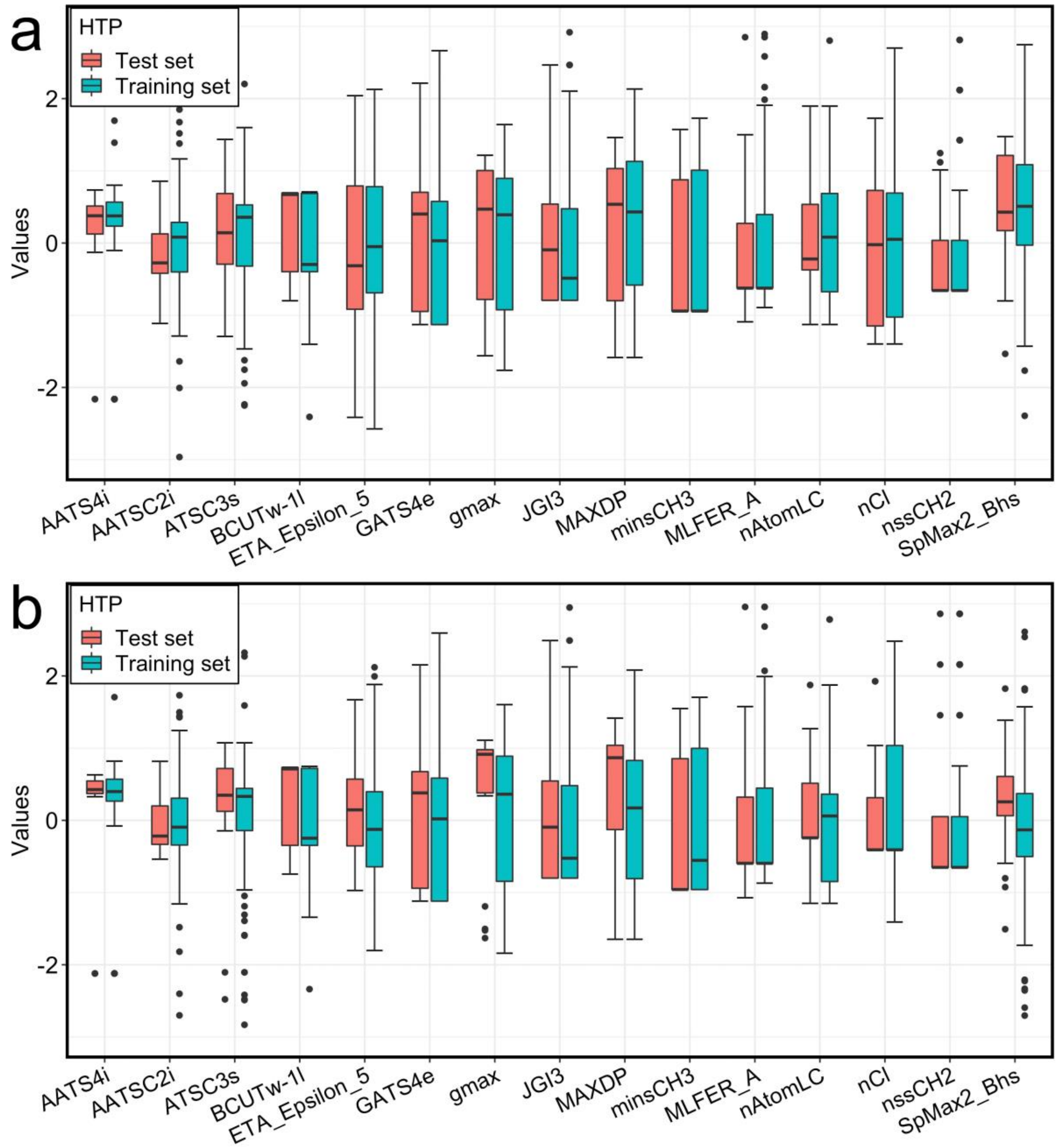

Figure S9. Ranges of the input feature values (after normalized) for training and test data points in HTP models under two scenarios: (a) a good predictive model (test set $R^2 = 0.70$) and (b) a poor model (test set $R^2 = 0.51$).

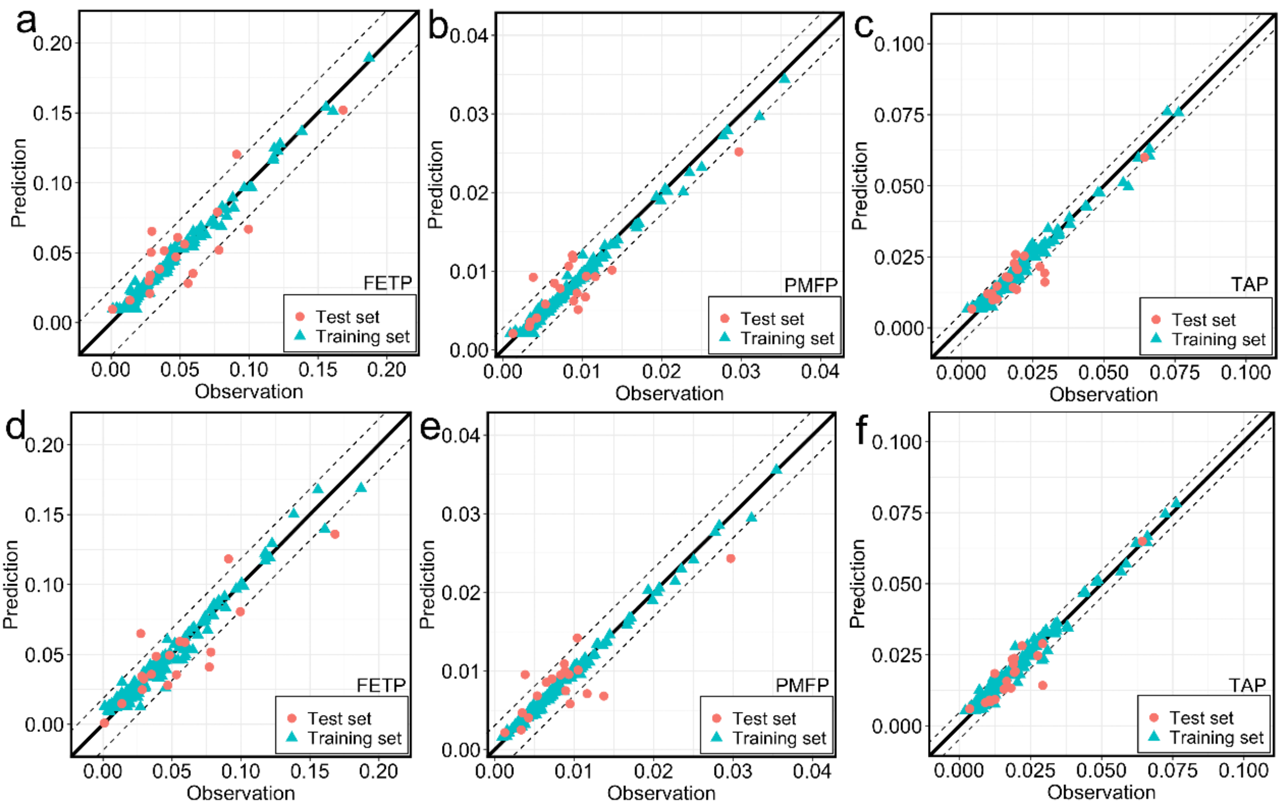

Figure S10. Prediction performance of MPN models (a-c) and improved MPEN models (d-f). The diagonal represents the perfect prediction line, and the dotted lines represent the intercepts of ±RMSE.

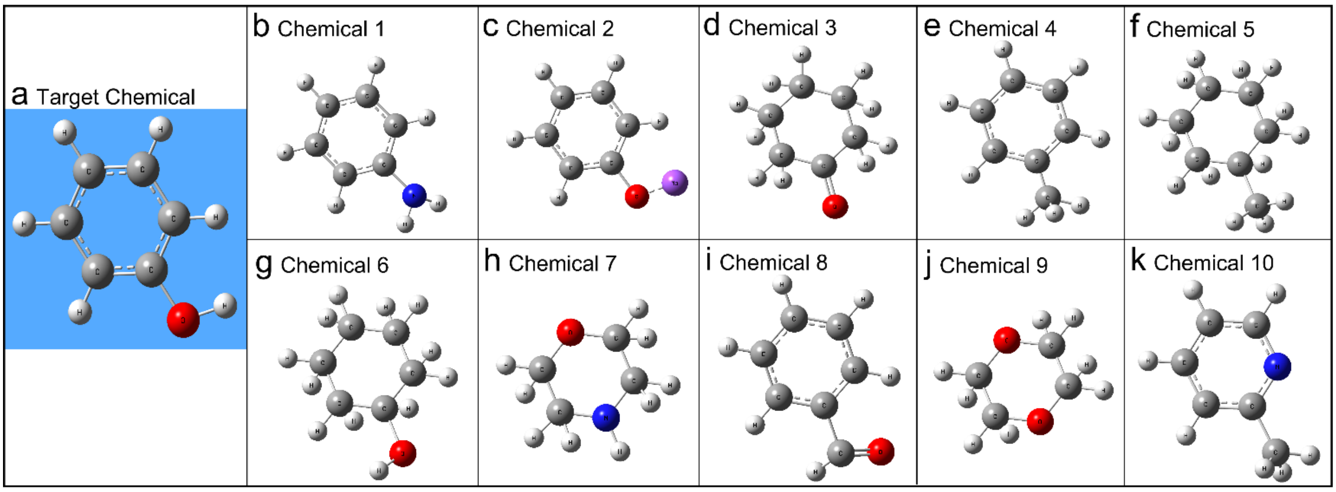

Figure S11. Structures of the target chemical (a) and the most relevant ten chemicals to target chemical (b-k). Phenol (a), aniline (b), sodium phenolate (c), cyclohexanone (d), toluene (e), methylcyclohexane (f), cyclohexanol (g), morpholine (h), benzaldehyde (i), dioxane (j) and alpha-picoline (k).

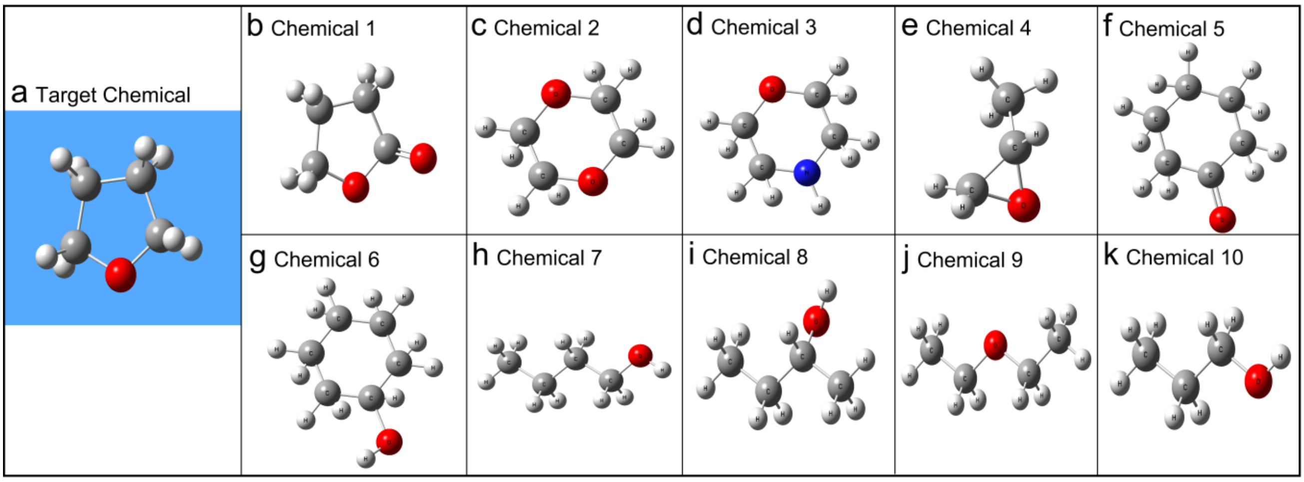

Figure S12. Structures of the target chemical (a) and the most relevant ten chemicals to target chemical (b-k). Tetrahydrofuran (a), butyrolactone (b), dioxane (c), morpholine (d), propylene oxide (e), cyclohexanone (f), cyclohexanol (g), 1-butanol (h), 2-butanol (i), diethyl ether (j) and 1-propanol (k).

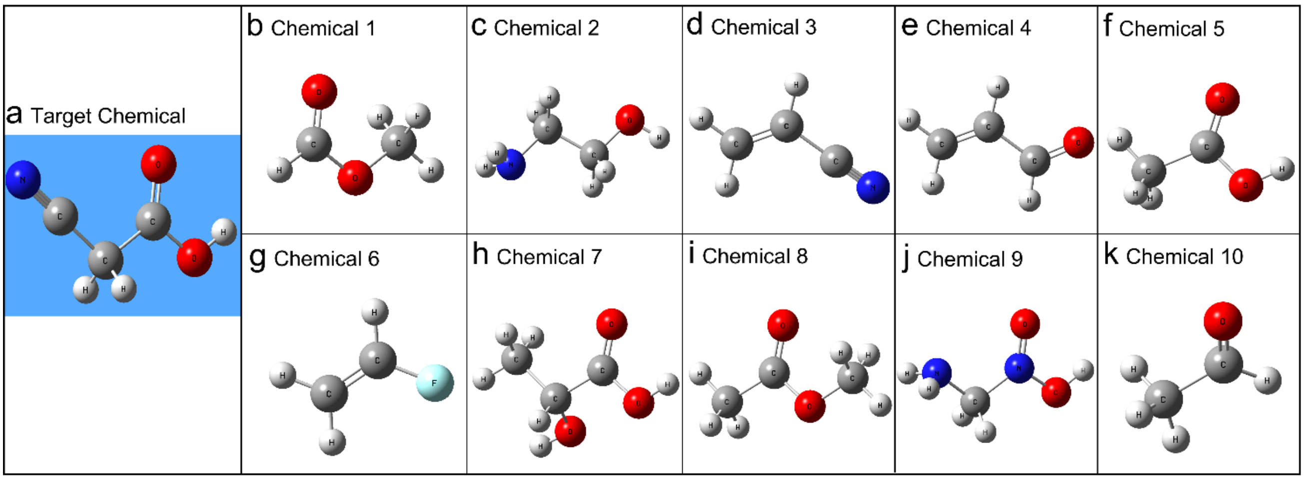

Figure S13. Structures of the target chemical (a) and the most relevant ten chemicals to target chemical (b-k). Cyanoacetic acid (a), methyl formate (b), monoethanolamine (c), acrylonitrile (d), acrolein (e), acetic acid (f), vinyl fluoride (g), lactic acid (h), methyl acetate (i), glycine (j) and acetaldehyde (k).

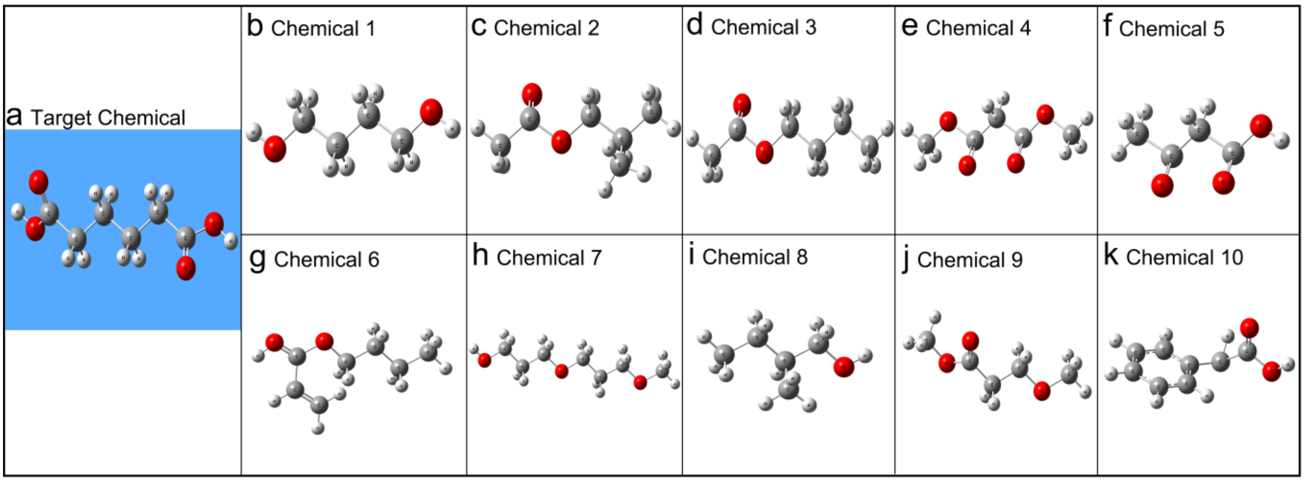

Figure S14. Structures of the target chemical (a) and the most relevant ten chemicals to target chemical (b-k). Adipic acid (a), butane-1, 4-diol (b), isobutyl acetate (c), butyl acetate (d), dimethyl malonate (e), acetoacetic acid (f), butyl acrylate (g), dipropylene glycol monomethyl ether (h), 2-methyl-1-butanol (i), methyl-3-methoxypropionate (j) and phenyl acetic acid (k).

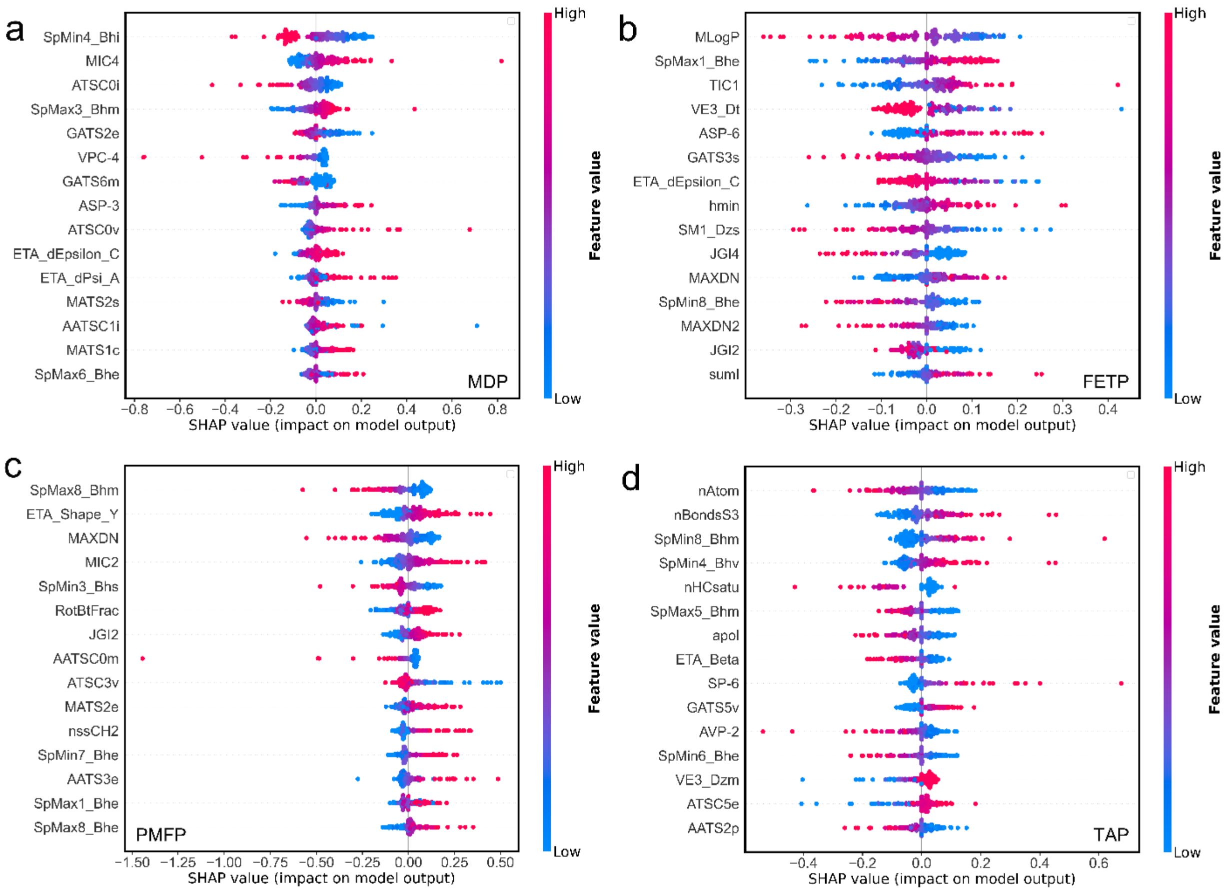

Figure S15. Importance of the representative descriptors (top 15) and Shapley values for (a) MDP, (b) FETP, (c) PMFP and (d) TAP MPEN models.

Table S1. The detailed data list of the collected chemicals, which includes the SMILES format and the environmental impact values (GWP, HTP, MDP, FETP, PMFP and TAP).

	Name	SMILEs	Environmental Impacts					
			GWP	HTP	MDP	FETP	PMFP	TAP
			kg CO2-Eq	kg 1,4-DCB-Eq	kg Fe-Eq	kg 1,4-DCB-Eq	kg PM10-Eq	kg SO2-Eq
1	1-propanol	CCCO	4.44E+00	1.56E+00	1.54E-01	4.07E-02	8.66E-03	2.04E-02
2	1,1-difluoroethane	CC(F)F	6.40E+00	2.68E+00	3.14E-01	7.39E-02	1.45E-02	2.93E-02
3	1-butanol	CCCCO	2.90E+00	8.21E-01	8.78E-02	2.19E-02	4.89E-03	1.19E-02
4	2-butanol	CCC(C)O	4.29E+00	7.60E-01	7.61E-02	1.76E-02	5.22E-03	1.45E-02
5	isobutanol	CC(C)CO	2.90E+00	8.21E-01	8.78E-02	2.19E-02	4.89E-03	1.19E-02
6	2-methyl-2-butanol	CCC(C)(C)O	3.16E+00	8.30E-01	7.23E-02	1.99E-02	5.20E-03	1.48E-02
7	2-nitroaniline	C1=CC=C(C(=C1)N)[N+](=O)[O-]	6.86E+00	2.53E+00	3.72E-01	1.21E-01	1.23E-02	2.81E-02
8	2, 4-dichlorophenol	C1=CC(=C(C=C1Cl)Cl)O	4.48E+00	2.34E+00	2.13E-01	5.25E-02	9.67E-03	1.78E-02
9	2, 4-dichlorotoluene	CC1=C(C=C(C=C1)Cl)Cl	3.33E+00	1.74E+00	1.52E-01	3.61E-02	6.70E-03	1.30E-02
10	3-methyl-1-butyl acetate	CC(C)CCOC(=O)C	5.38E+00	1.62E+00	1.58E-01	4.07E-02	9.14E-03	2.36E-02
11	4-methyl-2-pentanone	CC(C)CC(=O)C	4.28E+00	7.91E-01	9.22E-02	3.74E-02	6.77E-03	1.80E-02
12	4-tert-butylbenzaldehyde	CC(C)(C)C1=CC=C(C=C1)C=O	4.60E+00	1.20E+00	1.79E-01	3.26E-02	6.01E-03	1.35E-02
13	4-tert-butyltoluene	CC1=CC=C(C=C1)C(C)(C)C	2.37E+00	3.50E-01	5.77E-02	9.81E-03	2.60E-03	6.51E-03
14	acetaldehyde	CC=O	1.84E+00	4.63E-01	5.75E-02	1.49E-02	3.00E-03	6.83E-03
15	acetanilide	CC(=O)NC1=CC=CC=C1	5.48E+00	2.46E+00	3.57E-01	6.21E-02	1.09E-02	2.71E-02
16	acetic acid	CC(=O)O	1.72E+00	7.29E-01	9.86E-02	2.29E-02	3.57E-03	7.66E-03
17	acetic anhydride	CC(=O)OC(=O)C	3.53E+00	1.27E+00	1.70E-01	4.44E-02	5.46E-03	1.28E-02

18	acetoacetic acid	CC(=O)CC(=O)O	8.35E+00	2.62E+00	4.09E-01	7.82E-02	1.16E-02	2.75E-02
19	acetone	CC(=O)C	2.27E+00	2.02E-01	2.36E-02	1.70E-02	3.20E-03	9.35E-03
20	acetyl chloride	CC(=O)Cl	7.36E+00	2.58E+00	3.35E-01	8.39E-02	1.13E-02	2.51E-02
21	acetylene	C#C	6.46E+00	2.31E+00	1.03E-01	7.77E-02	1.69E-02	2.62E-02
22	acrolein	C=CC=O	2.43E+00	5.77E-01	6.10E-02	1.73E-02	3.63E-03	7.94E-03
23	acrylic acid	C=CC(=O)O	2.13E+00	3.44E-01	5.65E-02	9.37E-03	2.09E-03	4.66E-03
24	allyl chloride	C=CCCl	1.51E+00	6.79E-01	7.57E-02	1.68E-02	2.74E-03	6.04E-03
25	alpha-naphthol	C1=CC=C2C(=C1)C=CC=C2O	3.15E+00	1.67E+00	2.46E-01	4.27E-02	7.73E-03	2.01E-02
26	alpha-picoline	CC1=CC=CC=N1	3.86E+00	1.12E+00	1.54E-01	3.25E-02	6.27E-03	1.49E-02
27	aniline	C1=CC=C(C=C1)N	5.48E+00	2.46E+00	3.54E-01	5.62E-02	1.11E-02	2.89E-02
28	anthranilic acid	C1=CC=C(C(=C1)C(=O)O)N	8.96E+00	1.84E+00	5.05E-01	4.52E-02	1.05E-02	3.37E-02
29	benzal chloride	C1=CC=C(C=C1)C(Cl)Cl	2.75E+00	1.42E+00	1.18E-01	2.62E-02	4.81E-03	1.03E-02
30	benzaldehyde	C1=CC=C(C=C1)C=O	5.12E+00	2.57E+00	2.42E-01	4.82E-02	8.75E-03	1.93E-02
31	benzyl alcohol	C1=CC=C(C=C1)CO	4.34E+00	1.89E+00	2.34E-01	3.98E-02	7.27E-03	1.80E-02
32	benzyl chloride	C1=CC=C(C=C1)CCl	2.63E+00	1.04E+00	9.66E-02	2.01E-02	4.04E-03	9.12E-03
33	bisphenol A	CC(C)(C1=CC=C(C=C1)O)C2=CC=C(C=C2)O	4.12E+00	1.19E+00	1.46E-01	3.64E-02	7.21E-03	1.43E-02
34	bromopropane	CCCBr	5.16E+00	1.50E+00	1.70E-01	3.82E-02	7.51E-03	1.90E-02
35	butane	CCCC	8.20E-01	2.24E-01	2.74E-02	1.19E-02	1.53E-03	3.77E-03
36	butane-1, 4-diol	CC(C)(CO)CO.C(CCC(=O)O)CC(=O)O.C(CCO)CO	5.29E+00	1.81E+00	1.73E-01	5.32E-02	1.05E-02	2.20E-02
37	butyl acetate	CCCCOC(=O)C	3.73E+00	1.28E+00	1.65E-01	3.41E-02	6.60E-03	1.59E-02
38	butyl acrylate	CCCCOC(=O)C=C	4.27E+00	1.17E+00	1.40E-01	2.79E-02	7.09E-03	1.88E-02

39	carbon tetrachloride	C(Cl)(Cl)(Cl)Cl	1.70E+00	3.56E-01	6.53E-03	4.71E-03	4.63E-03	1.15E-02
40	chloroacetic acid	C(C(=O)O)Cl	2.35E+00	1.62E+00	1.68E-01	3.51E-02	5.34E-03	1.09E-02
41	chloroacetyl chloride	C(C(=O)Cl)Cl	4.55E+00	3.34E+00	3.81E-01	7.20E-02	1.41E-02	3.79E-02
42	chloromethyl methyl ether	COCCl	1.66E+00	1.10E+00	1.45E-01	2.70E-02	3.28E-03	7.10E-03
43	chloronitrobenzene	C1=CC=C(C(=C1)[N+](=O)[O-])Cl	4.64E+00	1.63E+00	2.43E-01	9.10E-02	8.28E-03	1.90E-02
44	chloropropionic acid	CC(C(=O)O)Cl	3.49E+00	1.89E+00	1.87E-01	4.34E-02	7.46E-03	1.47E-02
45	cumene	CC(C)C1=CC=CC=C1	2.51E+00	6.61E-01	6.11E-02	1.75E-02	4.72E-03	8.92E-03
46	cyanogen chloride	C(#N)Cl	5.25E+00	2.38E+00	2.08E-01	5.76E-02	9.61E-03	2.09E-02
47	cyanuric chloride	C1(=NC(=NC(=N1)Cl)Cl)Cl	5.70E+00	2.73E+00	2.65E-01	6.51E-02	1.06E-02	2.27E-02
48	cyclohexane	C1CCCCC1	2.60E+00	8.13E-01	6.65E-02	2.03E-02	5.74E-03	1.08E-02
49	cyclohexanol	C1CCC(CC1)O	2.94E+00	9.89E-01	1.08E-01	2.66E-02	5.91E-03	1.09E-02
50	cyclohexanone	C1CCC(=O)CC1	4.51E+00	1.43E+00	1.29E-01	3.85E-02	8.83E-03	1.93E-02
51	dichloromethane	C(Cl)Cl	3.48E+00	5.57E-01	5.44E-03	4.38E-03	9.26E-03	2.12E-02
52	diethanolamine	C(CO)NCCO	2.89E+00	8.71E-01	1.19E-01	2.34E-02	4.45E-03	1.02E-02
53	diethyl ether	CCOCC	5.20E+00	2.43E+00	3.59E-01	5.86E-02	1.15E-02	2.95E-02
54	diethylene glycol	C(COCCO)O	2.24E+00	7.31E-01	1.04E-01	2.03E-02	3.80E-03	7.78E-03
55	dimethyl ether	COC	1.42E+00	5.93E-01	1.02E-01	2.24E-02	2.19E-03	5.11E-03
56	dimethyl malonate	COC(=O)CC(=O)OC	5.73E+00	3.12E+00	4.10E-01	8.59E-02	1.11E-02	2.40E-02
57	dimethyl sulfate	COS(=O)(=O)OC	1.42E+00	1.10E+00	1.95E-01	2.90E-02	4.26E-03	1.25E-02
58	dimethyl sulfide	CSC	1.84E+00	7.77E-01	1.27E-01	2.71E-02	3.12E-03	7.44E-03
59	dimethyl sulfoxide	CS(=O)C	1.46E+00	6.69E-01	1.13E-01	2.29E-02	2.57E-03	5.96E-03
60	dimethylacetamide	CC(=O)N(C)C	3.25E+00	1.30E+00	1.99E-01	4.07E-02	5.91E-03	1.35E-02

61	dimethylamine	CNC	2.47E+00	8.11E-01	1.33E-01	2.80E-02	3.71E-03	1.01E-02
62	dioxane	C1COCCO1	4.89E+00	1.56E+00	2.01E-01	4.10E-02	7.99E-03	1.78E-02
63	dipropyl amine	CCCNCCC	5.69E+00	1.84E+00	1.87E-01	4.67E-02	1.02E-02	2.57E-02
64	dipropylene glycol monomethyl ether	CC(CO)OCC(C)OC	4.93E+00	2.81E+00	2.59E-01	6.04E-02	1.02E-02	1.95E-02
65	DTPA	C(CN(CC(=O)O)CC(=O)O)N(CCN(CC(=O)O)CC(=O)O)CC(=O)O	4.05E+00	1.54E+00	2.04E-01	6.41E-02	6.69E-03	1.57E-02
66	EDTA	C(CN(CC(=O)O)CC(=O)O)N(CC(=O)O)CC(=O)O	4.17E+00	1.57E+00	2.05E-01	6.52E-02	6.98E-03	1.61E-02
67	epichlorohydrin	C1C(O1)CCl	3.10E+00	1.33E+00	1.79E-01	3.42E-02	5.57E-03	1.12E-02
68	ethyl acetate	CCOC(=O)C	2.69E+00	9.77E-01	1.49E-01	2.60E-02	4.72E-03	1.15E-02
69	ethyl benzene	CCC1=CC=CC=C1	2.44E+00	6.97E-01	6.17E-02	1.83E-02	4.94E-03	9.15E-03
70	ethylamine	CCN	3.02E+00	8.18E-01	1.45E-01	1.96E-02	4.21E-03	1.17E-02
71	ethylene bromide	C(CBr)Br	6.47E+00	1.81E+00	1.99E-01	4.58E-02	8.92E-03	2.22E-02
72	ethylene carbonate	C1COC(=O)O1	1.62E+00	6.26E-01	1.08E-01	1.59E-02	2.47E-03	5.38E-03
73	ethylene dichloride	C(CCl)Cl	1.43E+00	7.59E-01	1.01E-01	1.85E-02	2.95E-03	5.92E-03
74	ethylene glycol diethyl ether	CCOCCOCC	3.56E+00	1.48E+00	2.17E-01	3.78E-02	6.35E-03	1.30E-02
75	ethylene glycol dimethyl ether	COCCOC	2.42E+00	8.81E-01	1.39E-01	2.73E-02	3.79E-03	8.19E-03
76	ethylene glycol monoethyl ether	CCOCCO	2.33E+00	7.05E-01	1.13E-01	1.86E-02	3.45E-03	7.67E-03
77	ethylene oxide	C1CO1	2.18E+00	4.93E-01	5.86E-02	1.46E-02	3.12E-03	6.76E-03
78	ethylenediamine	C(CN)N	5.29E+00	2.24E+00	2.59E-01	5.18E-02	9.66E-03	2.21E-02
79	formic acid	C(=O)O	2.50E+00	1.04E+00	1.05E-01	2.73E-02	3.45E-03	1.06E-02
80	glycerine	C(C(CO)O)O	2.10E+00	6.83E-01	1.09E-01	4.17E-02	7.06E-03	1.83E-02
81	glycine	C(C(=O)O)N	4.96E+00	2.98E+00	3.32E-01	8.00E-02	1.08E-02	2.29E-02

82	glyoxal	C(=O)C=O	2.99E+00	9.79E-01	1.07E-01	2.81E-02	5.11E-03	1.04E-02
83	hexafluoroethane	C(C(F)(F)F)(F)(F)F	1.06E+01	7.14E+00	7.70E-01	1.68E-01	2.97E-02	6.43E-02
84	hydroquinone	C1=CC(=CC=C1O)O	3.84E+00	1.33E+00	1.68E-01	4.24E-02	7.23E-03	1.40E-02
85	imidazole	C1=CN=CN1	4.93E+00	1.77E+00	2.26E-01	4.66E-02	8.12E-03	1.96E-02
86	isobutyl acetate	CC(C)COC(=O)C	3.83E+00	1.31E+00	1.69E-01	3.49E-02	6.70E-03	1.62E-02
87	isohexane	CCCC(C)C	9.79E-01	5.03E-01	1.04E-01	1.19E-02	2.15E-03	6.06E-03
88	isopropanol	CC(C)O	2.10E+00	4.14E-01	6.80E-02	1.08E-02	2.63E-03	7.07E-03
89	isopropyl acetate	CC(C)OC(=O)C	4.48E+00	1.52E+00	1.95E-01	4.12E-02	7.76E-03	1.85E-02
90	isopropylamine	CC(C)N	3.80E+00	9.39E-01	1.53E-01	2.38E-02	5.12E-03	1.40E-02
91	lactic acid	CC(C(=O)O)O	4.36E+00	1.53E+00	2.06E-01	4.32E-02	6.86E-03	1.68E-02
92	maleic anhydride	C1=CC(=O)OC1=O	2.50E+00	6.31E-01	6.36E-02	1.79E-02	2.80E-03	7.11E-03
93	melamine	C1(=NC(=NC(=N1)N)N)N	5.23E+00	1.87E+00	3.35E-01	4.53E-02	1.13E-02	3.36E-02
94	meta-phenylene diamine	C1=CC(=CC(=C1)N)N	2.15E+01	3.82E+00	5.16E-01	1.07E-01	2.74E-02	7.50E-02
95	methacrylic acid	CC(=C)C(=O)O	6.42E+00	2.14E+00	3.50E-01	6.88E-02	1.09E-02	3.09E-02
96	methane sulfonic acid	CS(=O)(=O)O	1.14E+00	1.07E+00	1.87E-01	2.50E-02	4.36E-03	1.38E-02
97	methanol	CO	3.15E-01	2.26E+00	4.13E-02	1.05E-02	9.27E-04	2.05E-03
98	methyl acrylate	COC(=O)C=C	2.73E+00	1.25E+00	1.17E-01	1.73E-02	2.71E-03	6.22E-03
99	methyl ethyl ketone	CCC(=O)C	1.83E+00	3.76E-01	6.05E-02	1.01E-02	2.39E-03	5.87E-03
100	methyl formate	COC=O	2.83E+00	1.16E+00	1.30E-01	3.82E-02	5.66E-03	1.24E-02
101	methyl iodide	CI	6.79E+00	1.86E+00	1.90E-01	5.43E-02	1.03E-02	2.34E-02
102	methyl tert-butyl ether	CC(C)(C)OC	1.13E+00	2.84E-01	6.40E-02	8.48E-03	1.48E-03	3.93E-03
103	methyl-3-methoxypropionate	COCCC(=O)OC	2.83E+00	1.32E+00	1.53E-01	3.62E-02	3.40E-03	7.42E-03

104	methylamine	CN	2.63E+00	8.32E-01	1.34E-01	2.59E-02	3.92E-03	1.16E-02
105	methylchloride	CCl	3.10E+00	3.81E-01	4.63E-03	3.57E-03	8.83E-03	1.97E-02
106	methylcyclohexane	CC1CCCCC1	3.65E+00	7.37E-01	7.59E-02	1.83E-02	5.28E-03	1.45E-02
107	N-methyl-2-pyrrolidone	CN1CCCC1=O	7.10E+00	2.66E+00	3.20E-01	7.71E-02	1.37E-02	2.92E-02
108	N, N-dimethylformamide	CN(C)C=O	2.84E+00	1.15E+00	1.68E-01	3.53E-02	5.22E-03	1.26E-02
109	naphthalene sulfonic acid	C1=CC=C2C(=C1)C=CC=C2S(=O)(=O)O	1.52E+00	1.18E+00	1.31E-01	2.89E-02	5.83E-03	1.04E-02
110	nitrobenzene	C1=CC=C(C=C1)[N+](=O)[O-]	3.57E+00	1.18E+00	2.23E-01	2.84E-02	7.20E-03	1.96E-02
111	o-aminophenol	C1=CC=C(C(=C1)N)O	6.31E+00	1.87E+00	3.55E-01	4.85E-02	1.00E-02	2.39E-02
112	o-chlorobenzaldehyde	C1=CC=C(C(=C1)C=O)Cl	8.96E+00	4.10E+00	4.67E-01	1.00E-01	1.68E-02	3.24E-02
113	o-chlorotoluene	CC1=CC=CC=C1Cl	2.91E+00	1.23E+00	1.18E-01	2.64E-02	5.18E-03	1.06E-02
114	o-cresol	CC1=CC=CC=C1O	3.92E+00	1.28E+00	1.58E-01	3.69E-02	7.06E-03	1.37E-02
115	o-nitrophenol	C1=CC=C(C(=C1)[N+](=O)[O-])O	4.07E+00	1.13E+00	2.23E-01	2.95E-02	6.48E-03	1.56E-02
116	ortho-phenylene diamine	C1=CC=C(C(=C1)N)N	1.52E+01	4.15E+00	5.77E-01	1.61E-01	2.27E-02	5.68E-02
117	p-chlorophenol	C1=CC(=CC=C1O)Cl	4.34E+00	1.97E+00	1.94E-01	4.69E-02	8.91E-03	1.66E-02
118	p-nitrophenol	C1=CC(=CC=C1[N+](=O)[O-])O	4.07E+00	1.13E+00	2.23E-01	2.95E-02	6.48E-03	1.56E-02
119	p-nitrotoluene	CC1=CC=C(C=C1)[N+](=O)[O-]	3.62E+00	5.54E-01	1.57E-01	1.36E-02	3.81E-03	1.24E-02
120	para-phenylene diamine	C1=CC(=CC=C1N)N	1.22E+01	4.31E+00	6.07E-01	1.87E-01	2.04E-02	4.79E-02
121	pentaerythritol	C(C(CO)(CO)CO)O	2.41E+00	9.87E-01	1.43E-01	2.95E-02	4.53E-03	1.00E-02
122	perfluoropentane	C(C(C(F)(F)F)(F)F)(C(C(F)(F)F)(F)F)(F)F	1.64E+01	3.73E+00	4.06E-01	8.05E-02	1.93E-02	4.36E-02
123	phenyl acetic acid	C1=CC=C(C=C1)CC(=O)O	5.78E+00	2.14E+00	2.59E-01	5.91E-02	8.98E-03	2.10E-02
124	phenyl isocyanate	C1=CC=C(C=C1)N=C=O	7.76E+00	5.15E+00	5.00E-01	9.65E-02	1.66E-02	3.82E-02
125	phosgene	C(=O)(Cl)Cl	1.38E+00	2.72E+00	1.23E-01	2.98E-02	2.62E-03	7.38E-03

126	phthalic anhydride	C1=CC=C2C(=C1)C(=O)OC2=O	2.61E+00	4.93E-01	6.06E-02	1.25E-02	4.29E-03	1.12E-02
127	phthalimide	C1=CC=C2C(=C1)C(=O)NC2=O	3.73E+00	9.30E-01	1.32E-01	2.33E-02	6.22E-03	1.64E-02
128	piperidine	C1CCNCC1	8.88E+00	2.67E+00	3.67E-01	7.40E-02	1.28E-02	3.04E-02
129	polyacrylamide	C=CC(=O)N	2.84E+00	6.96E-01	1.16E-01	1.74E-02	4.45E-03	1.49E-02
130	propanal	CCC=O	3.74E+00	1.17E+00	9.81E-02	3.00E-02	6.92E-03	1.73E-02
131	propionic acid	CCC(=O)O	2.02E+00	6.91E-01	8.38E-02	2.00E-02	3.71E-03	8.10E-03
132	propyl amine	CCCN	6.45E+00	2.24E+00	2.51E-01	5.75E-02	1.19E-02	2.90E-02
133	propylene	CC=C	1.44E+00	9.46E-03	6.68E-04	9.71E-04	1.25E-03	3.61E-03
134	propylene glycol	CC(CO)O	4.54E+00	2.67E+00	2.46E-01	5.58E-02	9.47E-03	1.81E-02
135	propylene oxide	CC1CO1	5.00E+00	2.84E+00	2.39E-01	5.97E-02	1.05E-02	2.01E-02
136	pyrazole	C1=CNN=C1	1.84E+01	2.03E+01	9.27E-01	2.01E-01	3.23E-02	6.60E-02
137	pyridine	C1=CC=NC=C1	8.17E+00	2.37E+00	3.17E-01	6.64E-02	1.17E-02	2.80E-02
138	sodium methoxide	C[O-].[Na+]	1.70E+00	8.67E-01	1.09E-01	2.51E-02	3.79E-03	7.15E-03
139	styrene	C=CC1=CC=CC=C1	3.11E+00	8.70E-01	6.71E-02	2.41E-02	6.30E-03	1.14E-02
140	tert-butyl amine	CC(C)(C)N	7.82E+00	2.70E+00	3.62E-01	7.58E-02	1.36E-02	3.40E-02
141	tetrachloroethylene	C(=C(Cl)Cl)(Cl)Cl	3.92E+00	7.63E-01	4.47E-03	4.57E-03	6.86E-03	1.47E-02
142	tetraethyl orthosilicate	CCO[Si](OCC)(OCC)OCC	5.20E+00	2.43E+00	2.20E-01	5.53E-02	9.93E-03	2.03E-02
143	tetrafluoroethane	C(C(F)(F)F)F	7.57E+00	4.96E+00	6.16E-01	1.18E-01	2.07E-02	4.41E-02
144	toluene	CC1=CC=CC=C1	1.55E+00	2.39E-02	3.78E-03	1.12E-03	1.35E-03	3.81E-03
145	trichloroacetic acid	C(=O)(C(Cl)(Cl)Cl)O	4.07E+00	2.75E+00	2.38E-01	5.85E-02	9.76E-03	1.81E-02
146	trichloroethylene	C(=C(Cl)Cl)Cl	4.28E+00	2.43E+00	2.35E-01	5.93E-02	1.04E-02	1.85E-02
147	trichloromethane	C(Cl)(Cl)Cl	3.53E+00	7.38E+00	1.16E-01	2.58E-02	3.17E-03	6.94E-03

148	trichloropropane	CCC(Cl)(Cl)Cl	2.97E+00	2.06E+00	1.94E-01	9.96E-02	9.31E-03	2.93E-02
149	triethyl amine	CCN(CC)CC	3.10E+00	8.19E-01	1.49E-01	1.97E-02	4.13E-03	1.08E-02
150	trifluoroacetic acid	C(=O)(C(F)(F)F)O	9.03E+00	6.23E+00	6.67E-01	1.38E-01	2.35E-02	4.87E-02
151	trifluoromethane	C(F)(F)F	8.24E+00	1.57E+01	5.50E-01	8.81E-02	1.30E-02	3.43E-02
152	trimesoyl chloride	C1=C(C=C(C=C1C(=O)Cl)C(=O)Cl)C(=O)Cl	8.78E+00	3.21E+00	3.23E-01	6.02E-02	2.50E-02	6.60E-02
153	trimethyl borate	B(OC)(OC)OC	2.52E+00	1.01E+00	1.71E-01	2.99E-02	6.93E-03	1.43E-02
154	vinyl acetate	CC(=O)OC=C	2.29E+00	8.83E-01	1.23E-01	2.64E-02	4.24E-03	8.98E-03
155	vinyl chloride	C=CCl	1.60E+00	1.71E-01	2.59E-03	3.97E-03	1.36E-03	3.96E-03
156	vinyl fluoride	C=CF	9.50E+00	3.24E+00	3.86E-01	8.90E-02	1.71E-02	3.43E-02
157	xylene	CC1=CC=C(C=C1)C	1.69E+00	3.03E-02	3.86E-03	1.36E-03	1.55E-03	4.47E-03
158	2,4-di-tert-butylphenol	CC(C)(C)C1=CC(=C(C=C1)O)C(C)(C)C	3.43E+00	9.04E-01	1.14E-01	2.57E-02	5.29E-03	1.03E-02
159	2,6-di-tert-butylphenol	CC(C)(C)C1=C(C(=CC=C1)C(C)(C)C)O	3.57E+00	9.31E-01	1.17E-01	2.65E-02	5.48E-03	1.06E-02
160	sodium phenolate	C1=CC=C(C=C1)[O-].[Na+]	3.93E+00	1.43E+00	1.68E-01	3.92E-02	7.68E-03	1.45E-02
161	dichloropropene	CC=C(Cl)Cl	4.08E+00	1.85E+00	2.03E-01	4.58E-02	7.23E-03	1.59E-02
162	dimethyldichlorosilane	C[Si](C)(Cl)Cl	6.18E+00	1.55E+00	7.51E-02	3.51E-02	1.43E-02	2.97E-02
163	monochlorobenzene	C1=CC=C(C=C1)Cl	3.05E+00	1.48E+00	1.33E-01	1.02E-01	7.47E-03	1.34E-02
164	monochloropentafluoroethane	C(C(F)(F)Cl)(F)(F)F	9.46E+00	7.41E+00	8.30E-01	1.56E-01	2.77E-02	6.21E-02
165	o-dichlorobenzene	C1=CC=C(C(=C1)Cl)Cl	2.84E+00	1.45E+00	1.30E-01	1.18E-01	7.02E-03	1.25E-02
166	p-dichlorobenzene	C1=CC(=CC=C1Cl)Cl	2.84E+00	1.45E+00	1.30E-01	1.18E-01	7.02E-03	1.25E-02
167	sodium chloroacetate	C(C(=O)[O-])Cl.[Na+]	3.46E+00	2.08E+00	2.64E-01	5.20E-02	7.21E-03	1.64E-02
168	benzene	C1=CC=CC=C1	2.03E+00	5.14E-01	9.68E-03	1.36E-02	4.75E-03	8.22E-03
169	ethanol	CCO	1.25E+00	2.67E-01	5.39E-02	6.20E-03	1.57E-03	4.14E-03

170	ethylene glycol	C(CO)O	1.98E+00	6.72E-01	9.72E-02	1.85E-02	3.43E-03	6.96E-03
171	glucose	C(C1C(C(C(C(O1)O)O)O)O)O	1.39E+00	7.00E-01	1.27E-01	2.40E-02	3.35E-03	1.05E-02
172	1-pentanol	CCCCCO	5.05E+00	1.27E+00	7.56E-02	3.07E-02	8.32E-03	2.22E-02
173	2-methyl-1-butanol	CCC(C)CO	5.05E+00	1.27E+00	7.56E-02	3.07E-02	8.32E-03	2.22E-02
174	acetonitrile	CC#N	4.02E+00	6.61E-01	1.11E-01	1.63E-02	5.65E-03	2.21E-02
175	acrylonitrile	C=CC#N	2.98E+00	4.92E-01	8.25E-02	1.21E-02	4.21E-03	1.64E-02
176	benzoic acid	C1=CC=C(C=C1)C(=O)O	2.22E+00	4.81E-01	6.42E-02	1.38E-02	3.25E-03	7.00E-03
177	butyrolactone	C1CC(=O)OC1	6.28E+00	2.28E+00	2.43E-01	6.59E-02	1.26E-02	2.61E-02
178	decabromodiphenyl ether	C1(=C(C(=C(C(=C1Br)Br)Br)Br)OC2=C(C(=C(C(=C2Br)Br)Br)Br)Br	1.35E+01	3.73E+00	4.27E-01	1.23E-01	2.83E-02	7.63E-02
179	dimethyl carbonate	COC(=O)OC	2.22E+00	9.40E-01	1.59E-01	2.58E-02	3.59E-03	7.59E-03
180	ethane	CC	8.15E-01	2.23E-01	2.74E-02	1.19E-02	1.52E-03	3.76E-03
181	hexamethyldisilazane	C[Si](C)(C)N[Si](C)(C)C	5.84E+00	1.96E+00	3.73E-01	5.95E-02	1.14E-02	2.49E-02
182	methyl acetate	CC(=O)OC	1.15E+00	4.30E-01	6.48E-02	1.07E-02	2.42E-03	6.72E-03
183	monoethanolamine	C(CO)N	2.87E+00	8.64E-01	1.21E-01	2.28E-02	4.44E-03	1.08E-02
184	morpholine	C1COCCN1	9.53E+00	2.73E+00	2.35E-01	8.33E-02	1.99E-02	5.86E-02
185	polydimethylsiloxane	C[Si](C)(C)O[Si](C)(C)O[Si](C)(C)C	1.55E+01	4.77E+00	3.64E-01	1.18E-01	3.55E-02	7.23E-02
186	propane	CCC	8.28E-01	1.39E-01	1.73E-02	4.46E-03	1.78E-03	5.62E-03
187	salicylic acid	C1=CC=C(C(=C1)C(=O)O)O	4.90E+00	2.32E+00	3.27E-01	5.96E-02	1.10E-02	2.59E-02

Table S2. The mean, minimum, median, maximum and standard deviation of each impact category.

	GWP	HTP	MDP	FETP	PMFP	TAP
	kg CO_2-Eq	kg 1,4-DCB-Eq	kg Fe-Eq	kg 1,4-DCB-Eq	kg PM_{10}-Eq	kg SO_2-Eq
Mean	4.36E+00	1.79E+00	1.94E-01	4.42E-02	8.02E-03	1.86E-02
Median	3.57E+00	1.27E+00	1.53E-01	3.42E-02	6.70E-03	1.47E-02
Minimum	3.15E-01	9.46E-03	6.68E-04	9.71E-04	9.27E-04	2.05E-03
Maximum	2.51E+01	2.03E+01	9.27E-01	2.01E-01	3.55E-02	7.63E-02
Standard deviation	3.22E+00	2.12E+00	1.53E-01	3.55E-02	6.03E-03	1.41E-02

Table S3. The best hyper-parameters optimized according to the Grid search for each model.

Impact categories	Learning rate	Hidden layers	Hidden neurons	Activation function	Optimizer function	Init mode	Batch size
GWP	0.001	2	64	LeakyReLU	Adam	he_uniform	128
FETP	0.001	1	128	LeakyReLU	Adam	he_uniform	128
MDP	0.001	1	128	LeakyReLU	Adam	he_uniform	64
TAP	0.0001	2	128	LeakyReLU	Adam	he_uniform	128
HTP	0.001	1	128	LeakyReLU	Adam	he_uniform	128
PMFP	0.001	2	64	LeakyReLU	Adam	he_uniform	128

Table S4. Performances of different machine learning algorithms for each impact categories.

Impact categories	RF		SVM		XGBoost		ANN	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
GWP	0.56	2.14	0.35	3.34	0.55	2.05	0.59	1.89
FETP	0.45	0.027	0.23	0.101	0.51	0.029	0.55	0.032
MDP	0.52	0.082	0.48	0.231	0.57	0.943	0.58	0.072
TAP	0.30	0.0067	0.08	0.0173	0.25	0.072	0.57	0.0061
HTP	0.61	1.01	0.41	2.08	0.62	1.05	0.56	1.06
PMFP	0.44	0.0032	0.15	0.0082	0.50	0.0035	0.51	0.0036

Table S5. Performance (RMSE) 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.

Impact categories		Data dimensionality reduction approaches				
		FULL	PCA	MI	PI	MIPI
RMSE	GWP	1.6192	1.3840	1.4131	1.4445	1.3563
	HTP	1.1874	1.4408	1.4588	1.1197	1.1055
	MDP	0.0900	0.0851	0.0955	0.0880	0.0865
	FETP	0.0220	0.0210	0.0210	0.0200	0.0205
	PMFP	0.0040	0.0035	0.0036	0.0035	0.0027
	TAP	0.0074	0.0068	0.0069	0.0062	0.0051

Table S6. Performances (R^2 and RMSE) of the models developed by input features based on three different similarity methods.

Impact categories	Cosine similarity		Euclidean distance		Weighted Euclidean distance	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
GWP	0.7503	0.9931	0.7431	1.1133	0.8071	0.9852
HTP	0.7731	0.6752	0.7723	0.6855	0.8091	0.6295
MDP	0.7946	0.0801	0.8016	0.0712	0.8419	0.0602
FETP	0.7315	0.0189	0.7132	0.0205	0.7524	0.0184
PMFP	0.6921	0.0044	0.6553	0.0051	0.7252	0.0031
TAP	0.8261	0.0056	0.7969	0.0069	0.8559	0.0047

References

(1) 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.

(2) 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

(3) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J. J.; Zhang, H. Machine learning: new ideas and tools in environmental science and engineering. *Environ. Sci. Technol.* **2021,** *55*, 12741–12754.

(4) 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.

(5) 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.

(6) 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.

(7) Lesnik, K. L.; Liu, H. Predicting microbial fuel cell biofilm communities and bioreactor performance using artificial neural networks. *Environ. Sci. Technol.* **2017,** *51,* 10881–10892.

(8) 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.

(9) Mao, J.; Jain, A. K. Artificial neural networks for feature extraction and multivariate data projection. *IEEE T. Neural Networ.* **2002,** *6*, 296–317.

(10) Wang, Y. S.; Linghu, R. K.; Zhang, W.; Shao, Y. C.; Xu, J. Study on deformation behavior in supercooled liquid region of a Ti-based metallic glassy matrix composite by artificial neural network. *J. Alloys Compd.* **2020,** *844*, 155761–155769.

(11) Hueffel. J. A.; Sperger. T.; Funes-Ardoiz. I.; Ward. J. S.; Rissanen. K.; Schoenebeck. F. Accelerated dinuclear palladium catalyst identification through unsupervised machine learning. *Science* **2021,** *374*, 1134–1140.

(12) Effrosynidis, D.; Arampatzis, A. An evaluation of feature selection methods for environmental data. *Ecol. Inform.* **2021,** *61*, 101224–101234.

(13) Altmann, A.; Tolosi, L.; Sander, O.; Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **2010,** *26*, 1340–1347.

(14) 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.

(15) Yuan, X.; Suvarna, M.; Low, S.; Dissanayake, P. D.; Lee, K. B.; Li, J.; Wang, X.; Ok, Y. S. Applied machine learning for prediction of CO_2 adsorption on biomass waste-derived porous carbons. *Environ. Sci. Technol.* **2021,** *55*, 11925–11936.

(16) 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.

(17) 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.