Environmental Science & Technology

pubs.acs.org/est

Article

Classification and Regression Machine Learning Models for Predicting Aerobic Ready and Inherent Biodegradation of Organic Chemicals in Water

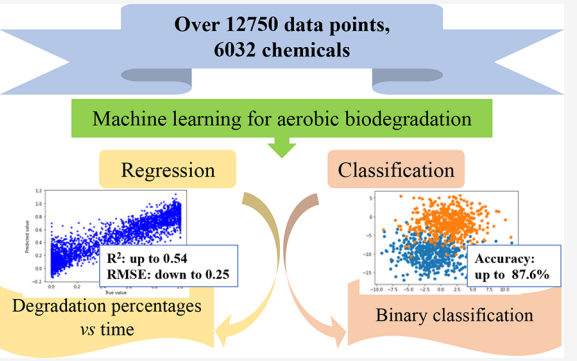
Kuan Huang and Huichun Zhang*

Cite This: https://doi.org/10.1021/acs.est.2c01764

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine learning (ML) is viewed as a promising tool for the prediction of aerobic biodegradation, one of the most important elimination pathways of organic chemicals from the environment. However, available models only have small datasets (<3200 records), make binary classification predictions, evaluate ready biodegradability, and do not incorporate experimental conditions (e.g., system setup and reaction time). This study addressed all these limitations by first compiling a large database of 12,750 records, considering both ready and inherent biodegradation under different conditions, and then developing regression and classification models using different chemical representations and ML algorithms. The best regression model (R^2 = 0.54 and root mean square error of 0.25) and classification model (the prediction accuracy from 85.1%) achieved very good performance. The model interpretation indicated that the models correctly captured the effects of chemical substructures, following the order of C═O > O═C−O > OH > CH_3 > halogen > branching > N > 6-member ring. The consideration of chemical speciation based on pK_a and α notations did not affect the regression model performance but significantly improved the classification model performance (the accuracy increased to 87.6%). The models also showed large applicability domains and provided reasonable predictions for more than 98% of over 850,000 environmentally relevant chemicals in the Distributed Structure-Searchable Toxicity database. These robust, trustable models were finally made widely accessible through two free online predictors with graphical user interface.



KEYWORDS: *closed bottle test, closed respirometer, CO_2 evolution test, DOC die away, EU method C.4, inherent biodegradation, OECD 301, ready biodegradation*

INTRODUCTION

Persistent, bioaccumulative, and toxic (PBT) organic chemicals are posing increasing risks to the environment. Among different transformation pathways, aerobic biodegradation is one of the key processes for the elimination of organic chemicals from the environment. To develop environmentally friendly chemicals and conduct accurate chemical risk assessment, quantifying the aerobic biodegradability of PBT chemicals is one of the most crucial tasks.[1]

Many organizations have issued standard methods for aerobic biodegradation tests, such as the International Organization for Standardization (ISO), Organization for Economic Cooperation and Development (OECD), American Society for Testing and Materials (ASTM International), the European Union (EU), the Japanese Ministry of International Trade and Industry (MITI), the National Institute of Technology and Evaluation (NITE), and the United States Environmental Protection Agency (EPA). Details about these methods can be found elsewhere.[2,3] These methods are mostly similar in terms of the scope, experimental setup, and applicability. As most experimental tests normally last at least 28 days, the cost for testing a sample commonly falls in the range of 3000−8000 USD or more in the United States depending on the methods (based on the communication with the industry). This is a significant financial burden to industries that need to develop environmentally friendly chemical products. Therefore, the use of quantitative structure−biodegradability relationships or machine learning (ML) approaches for biodegradability prediction has been encouraged.[1,4−6]

A number of studies have reported biodegradation prediction in the past few decades.[1,4,6−24] Among them, traditional statistical techniques such as linear, nonlinear, and partial least-squares regressions have been used in approaches

Received: March 11, 2022
Revised: June 27, 2022
Accepted: July 20, 2022

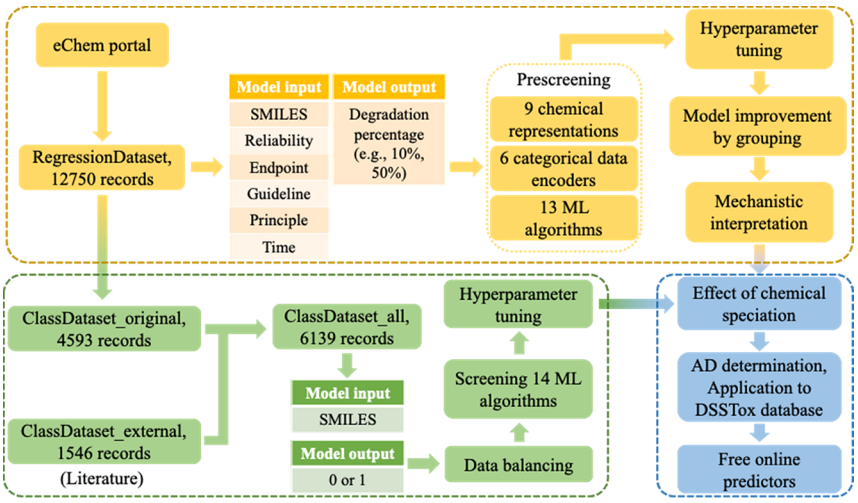


Figure 1. Flowchart of this study. Regression (yellow) and classification (green) models were built independently. The mechanistic interpretation was only performed on the regression model, while the last three steps (blue) , including investigating the effect of chemical speciation, determining AD, and developing free online predictors, were performed for both models.

such as the discriminant-function analysis, group contribution methods, and chemometric methods.[7,25−28] However, the development of such models usually requires comprehensive knowledge of the biodegradation mechanisms so that the most relevant physicochemical properties and/or substructures can be selected as the model input descriptors. As a result, a large number of models have been built on only narrowly defined, specific classes of chemicals with similar structures/properties. This significantly limits the applicability of these models.[28] With the rapid development of artificial intelligence in recent years, ML has become a popular option because it requires less domain knowledge and can handle much larger datasets regardless of whether the chemicals belong to one chemical class or not. For example, Cheng et al. developed combinatorial classification models using a support vector machine (SVM), *k*-nearest neighbors (KNN), naive Bayes, and C4.5 decision tree (DT) based on 1440 ready biodegradation datapoints using the MITI protocol.[1] The overall prediction accuracy for an external test dataset of 164 compounds was higher than 80%. Mansouri et al. developed multiple ready biodegradation prediction models using KNN, partial least-squares discriminant analysis, SVM, and their consensus models based on 1055 chemicals collected from NITE.[4] An external validation set consisting of 670 chemicals demonstrated sensitivity and specificity as high as 0.81 and 0.94, respectively. A recent study gathered all previously published ready biodegradation data for 2830 compounds from the literature and combined them with a set of 316 industrial chemicals.[24] Classification models were then built based on the largest dataset that we are aware of using SVM, random forest (RF), and naive Bayes and yielded balanced accuracies of 0.72−0.75. In addition, a few computer programs have been published in the past few years, such as BioWin, MultiCASE, CATABOL, VEGA, OPERA, and ToxTree, all of which incorporated biodegradation models and are therefore able to make biodegradability predictions. Most of them are widely recognized by both industries and academia. More information about these tools and other ML studies can be found in reviews.[28−32]

However, these models share four major limitations: (1) containing small numbers of chemicals (from <100 to 3200), (2) mostly built for binary classification purposes, that is, the outputs of the models are only 0s (not readily biodegradable) and 1s (readily biodegradable), (3) only capable of predicting ready but not inherent biodegradation, and (4) not considering the effects of experimental conditions, reaction time, and chemical speciation on biodegradability.[1,24]

To overcome the above limitations, we first compiled a comprehensive dataset consisting of 12,750 records for 6032 chemicals. The dataset contained six inputs—chemical SMILES strings, reaction time (day), guidelines (e.g., OECD 301D), principles (to describe system setup or condition, such as "closed bottle test" and "DOC die away"), endpoints (ready or inherent), and reliability levels (1 or 2)—and one output—the biodegradation percentages. To develop the best regression model, we compared the performance of 13 ML algorithms, 9 chemical representations, and 6 categorical encoding methods. The obtained models were carefully interpreted by evaluating the importance of the chemical substructures and all other input features using the SHapley Additive exPlanations (SHAP) method.[33] To expand the applicability of this study and compare it with the published models, we further built classification models by evaluating the performance of 14 ML algorithms and one chemical representation (the one used in the final regression model). The applicability domains (ADs) of the regression and classification models were defined and then applied to a large database called Distributed Structure-Searchable Toxicity (DSSTox), which is operated by the US EPA and currently contains over 850,000 environmentally relevant chemicals. Based on the obtained models, two user-friendly online predictors were made freely available at https://www.chemai.aropha.com/. The workflow of this study can be found in Figure 1. For those who have little ML knowledge, we recommend these reference papers,[34,35] which have step-by-step guidance for building ML models for general environmental applications.

METHODS

Dataset Preparation. The experimental data were mainly collected using the eChem portal (https://www.echemportal.org/echemportal/property-search), a tool that has access to a number of databases. Only the experimental data with

https://doi.org/10.1021/acs.est.2c01764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

reliability levels 1 (reliable without restriction) and 2 (reliable with restrictions) were retrieved to ensure high data quality.[24] Such restrictions may differ from one datapoint to another but can be looked up by entering the original webpage where the record is located. (In the original spreadsheet downloaded from the eChem portal, each record comes with a hyperlink. By clicking on this link, one will be redirected to the original webpage of that record.) All ready and inherent biodegradation data for aerobic conditions were included. This resulted in a total of more than 20,000 records. After curation, a dataset named "RegressionDataset" containing 12,750 data points was obtained for developing regression models. More details of the data curation process can be found in Text S1.1. A dataset called "ClassDataset_original" containing 4593 data points were obtained for classification models on the basis of the RegressionDataset. An external validation dataset containing 1546 chemicals, named "ClassDataset_external", was compiled from the literature because it would be meaningful to validate our models using data from a different source. Last, these two classification datasets were combined to form "ClassDataset_all," which was later used to build a comprehensive classification model. More details of the three datasets for classification models can be found in Text S1.2.

Selection of Optimum Chemical Representation, Categorical Data Encoder, and ML Algorithm. The type of chemical representation is one of the most important considerations in ML modeling as it transforms the raw chemical structural data into machine-readable information. An appropriate representation should cover all relevant features of the chemicals. To develop regression models, we compared seven molecular fingerprints (FPs, Table S1)—Atom pair, Topological torsion, Morgan, Pattern, RDK, MACCS, and PubChem, some with different dimensions—, one set of molecular descriptors (MDs), and a combination of MACCS and MDs. The model performance was evaluated based on root mean square errors (RMSE) and R^2 values. More details can be found in Text S1.3.

In addition to chemicals, the regression models included other variables/features, including reliability level, endpoint, guideline, principle, and reaction time as the inputs. More details of these features can be found in Tables S2 and S3. The features including endpoint, guideline, and principle are called categorical data and need to be encoded into numbers before model training. In this study, we compared six categorical encoders from the library category_encoders, namely, BinaryEncoder(), HashingEnoder(), OneHotEncoder(), OrdinalEncoder(), PolynomialEncoder(), and SumEncoder(), to examine their impact on the model performance. More details can be found in Text S1.4.

A total of 13 common ML algorithms (Table S4) were screened during the regression model development, including adaptive boosting (AdaB), bagging, DT, a deep neural network, extra trees, gradient boosting (GradientB), KNN, Lasso, linear regression (Lr), RF, Ridge, support vector regression, and XGBoost. As there are numerous research articles, official documentations, posts, forums, and tutorials about these algorithms,[35] we decided not to give introductions to them here. Note that all of these algorithms (as well as the chemical representations) were selected based on the literature as they were commonly used by other studies for environmental engineering applications.[33] However, our study is by far the most comprehensive in term of the number of algorithms and chemical representations considered. For example, among some of the best papers in this area, the numbers of evaluated algorithms and chemical representations were 4 and 8, 3 and 1, or 3 and 1.[1,4,24]

More details on how we screened the chemical representations, categorical data encoders, and ML algorithms when developing regression models can be found in Text S1.5.

To develop the classification models, the chemical representation selected for the final regression model (MACCS FPs) was used, and a total of 14 classification algorithms (Table S5) were compared, including AdaB, bagging, DT, ET, Gaussian Process, GradientB, linear SVM, Naive Bayes, Nearest Neighbors, Neural Net (Multi-layer Perceptron classifier), quadratic discriminant analysis, radial basis function SVM, RF, and XGBoost. The three datasets mentioned above (i.e., ClassDataset_original, ClassDataset_external, and ClassDataset_all) were used. The model performance was evaluated based on different matrices, including accuracy, sensitivity, specificity, balanced accuracy, ROC AUC, and f_1. More details can be found in Text S1.6.

Bayesian Optimization, Chemical Similarity Calculation and Grouping, and Feature Importance Evaluation by SHAP Analysis. The models developed in this study were further improved by performing Bayesian optimization for hyperparameter tuning. Chemical similarity calculation based on FPs and the Tanimoto index was then carried out for chemical grouping to ensure high similarities of chemicals among the training, validation, and test subsets.[36] To evaluate the importance and contribution of different features to biodegradation, we employed the SHAP method.[36] More details about these processes can be found in Texts S1.7−S1.9.

Individual Ready and Inherent Models, and Knowledge Transfer. To understand whether building a ready biodegradation model based on the ready biodegradation data and an inherent biodegradation model based on the inherent biodegradation data alone could give better prediction accuracy for either endpoint, we split the original dataset into ready and inherent biodegradation subsets and built a ready and an inherent biodegradation model separately. Then, we applied a technique called "knowledge transfer"[37] aiming at improving the performance of the inherent biodegradation model using the knowledge learned from the ready biodegradation model. To achieve that, we first used the ready biodegradation model to make predictions for the ready biodegradability of the chemicals in the inherent biodegradation dataset and then used the predictions as an additional input feature in the inherent biodegradation dataset to build an inherent biodegradation model. More details on how we performed knowledge transfer can be found in Text S1.10 and Figure S1.

Effect of Chemical Speciation on Biodegradability. When investigating the effects of chemical speciation on biodegradation, we did not select the dominant acid/base species under the experimental pH conditions because (1) there were 6032 chemicals so it was labor-intensive and (2) many chemicals had more than one major acid/base species in the solution. Instead, we used pK_a values and/or the α notations as part of the input features to capture chemical acid/base behaviors because these values were directly related to the speciation. As most of the chemicals in this study have unknown pK_a values, we used a high-quality online predictor to predict pK_a.[38] The predictor was built on more than 1.6 million chemicals and could identify up to 144 ionizable groups (acid or base) with a high prediction accuracy (R^2 =

https://doi.org/10.1021/acs.est.2c01764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

0.94−0.97 and RMSEs = 0.45−0.81). More details on how we obtained the pK_a values and calculated the corresponding α notations can be found in Text S1.11.

Model AD and the DSSTox Database. Model AD is used to evaluate if a model prediction is reliable for a query chemical. It can be determined by calculating the similarity between the chemical and every chemical in the training dataset of the model based on the Tanimoto index (eq S7). The maximum value among the obtained similarity values was used as the similarity between the query chemical and the dataset.[36] If this value is higher than the threshold specified for a model, the chemical is then considered as within the AD, or the prediction is reliable, and vice versa. In this study, we determined the ADs for both the regression and classification models and then applied them to the DSSTox database to examine the data coverage of these models. More details can be found in Text S1.12.

RESULTS AND DISCUSSION

Datasets for Regression and Classification Models. In the assembled regression dataset, the input features included chemical structures described by SMILES strings, reliability level, endpoint, principle, guideline, and reaction time. The data distribution shows that 71.1% of the 12,750 records have a reliability score of 1 (reliable without restrictions), and 90.0% are ready biodegradation data. The closed respirometer (42.2% of the records) is the most popular principle, followed by CO_2 evolution (22.1%), closed bottle test (20.0%), and DOC die away (15.7%). Among the guidelines, the OECD methods including 301B, C, D, and F are the most popular, contributing to 75.6% of the total records. A summary of the corresponding principles and endpoints for different guidelines can be found in Table S3. Figure S2 shows the number of chemicals having different biodegradation percentages.

For the three classification datasets, ClassDataset_original (4593 records), ClassDataset_external (1546 records), and ClassDataset_all (6139 records) all have around 65% of negative (0s, NRB) and 35% of positive (1s, RB) records (details in Text S1.2 and Table S6).

Prescreening of ML Algorithms, Categorical Encoders, and Chemical representations, and Performing Bayesian Optimization for Regression Models. A prescreening process was performed for 13 ML algorithms, 6 categorical encoders, and 9 chemical representations when developing the regression models, and the results are shown in Figures S3−S6. As it was time consuming and labor intensive to tune the hyperparameters of all algorithms under different conditions (categorical encoders and chemical representations), at this stage, we only used the default hyperparameters for the prescreening purpose. Overall, XGBoost was found to be the best algorithm, while ordinal encoder and MACCS FPs were the best categorical encoder and chemical representation, respectively. To better understand how MACCS bits represent chemical structures, we highlighted each bit (substructure) for phenol, as an example, in Table S7. More results of this prescreening process can be found in Text S2.1.

After conducting Bayesian optimization (Figure S7 and Table S8), the regression model achieved an R^2 of 0.54 and an RMSE of 0.25 (equivalent to 25% of biodegradation). Although the RMSE value is still high considering the degradation percentage range of 0−100%, this error is mainly due to the limitations of the standard biodegradation test guidelines, such as high variability among different tests, within the same lab or among different labs, for inoculum collected at different times and/or locations, and using different analytical methods.[39] Therefore, to improve the ML model performance, future studies are warranted to improve the data quality.

Although this performance may not look as good as those in some other studies, those studies are based on small numbers of narrowly defined, specific classes of chemicals (e.g., substituted benzenes), on which the models can easily be built to achieve higher prediction accuracies.[22,23,40] However, those reported models typically have much smaller ADs and can only be applied to limited numbers of chemicals. Nevertheless, the medium R^2 value of the best regression model is likely because there is not enough chemical diversity in the dataset to capture all structure−biodegradation relationships, despite that >6000 chemicals were incorporated already (see more discussion below). Future work is warranted to increase the number of diverse chemicals in the dataset to further improve the model performance.

Training with Different Sample Sizes to Improve the Model Performance. The above best regression model was based on the entire dataset, but it is of great interest to know whether changing the sample size of the training dataset can improve the model performance or not (for the prediction accuracy on the same test set). This is important because ML models often benefit from big data, but in some cases adding more data may introduce noise and result in worse model performance. Toward this goal, we first investigated how the similarity among the training, validation, and test datasets impacts the model performance. This was done by building a series of models using modified sub-datasets that contained fractions of chemicals whose similarity was above certain scores, for example, >0.9, >0.8... More details can be found in Text S2.2. The results suggested that higher similarities among the training, validation, and test datasets always gave better model performance (Figure S8), in agreement with other studies.[36,41]

For a model built above, while keeping the validation and test datasets unchanged (e.g., each having 10% of the 5000 records) (Figure S9A), we combined the other 80% of the 5000 records with the "remaining data" (7750) from the pool of 12,750 records to form an "enhanced training" set and examined whether this can improve the model performance on the same test set. As shown in Figures S9B and S10, the models using the enhanced training sets generally outperformed those using the original training sets. This is likely due to the relatively high similarity between the remaining data and the test set. As the "5000" model was taken as an example, 96.3% of the remaining 7750 data points were observed to have similarity scores higher than 0.5 when compared to the test subset in the "5000" model (Figure S11). The inclusion of these additional chemicals in the training dataset hence should have provided more structure-biodegradability information to improve the model performance. This also indicates that there is no need to build individual models based on sub-datasets of chemicals in order to achieve better model performance for a query chemical in real applications. Therefore, the regression model built above on the whole dataset is used for the rest of the study unless otherwise noted.

Mechanistic Interpretation (Feature Importance). To identify the most influential factors on biodegradation, we calculated the SHAP values for all 171 features, including 166 MACCS FPs, reaction time, guideline, principle, endpoint, and reliability. As the model development process involved a few

https://doi.org/10.1021/acs.est.2c01764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

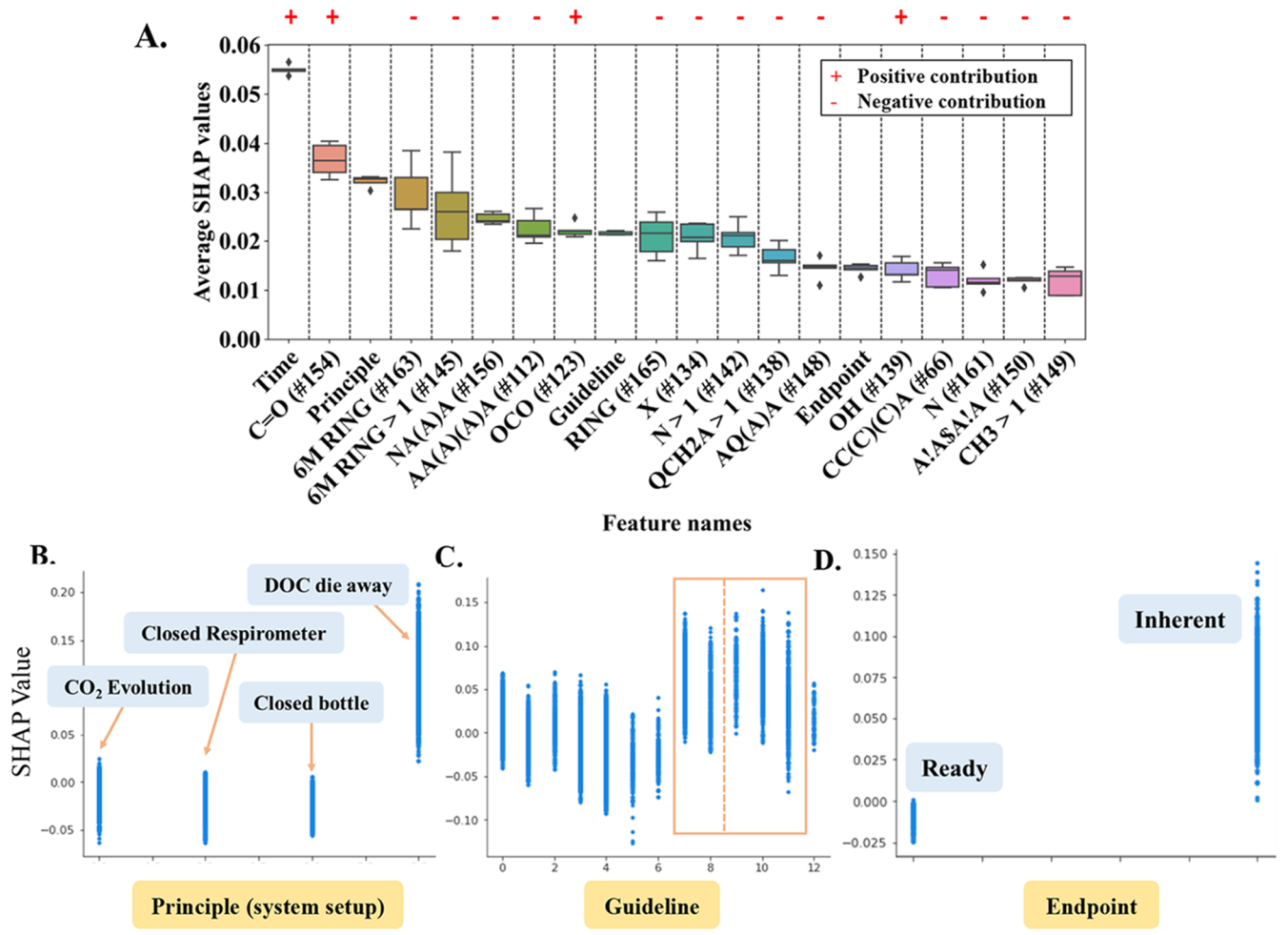


Figure 2. (A) Top 20 most important features based on their mean absolute SHAP values. The box plots show their importance scores while the red "+" and "−" symbols above the figure indicate positive and negative contributions to the biodegradation, respectively. The numbers following some of the feature names (all are substructures in MACCS) are their original orders in the MACCS naming system. A: any valid periodic table element symbol; Q: hetro atoms, any non-C or non-H atom; X: halogens; = : double bond; $: ring bond; !: chain or nonring bond. The SHAP values for different (B) principles, (C) guidelines, and (D) endpoints. Greater SHAP values suggest larger contributions to the biodegradation. The guidelines in (C) from left to right (0−12) are OECD 301B (0), OECD 301F (1), OECD 310 (2), OECD 301D (3), OECD 301C (4), EU C.4-D (5), EU C.4-E (6), OECD 302B (7), OECD 302C (8), EU C.4-A (9), OECD 301A (10), OECD 301E (11), and EU C.4-C (12). In (C), the left red box indicates the two inherent biodegradation guidelines, while the right red box includes the three ready biodegradation guidelines under the "DOC die away" system setup.

times of random splitting of the data, each experiment may result in slightly different SHAP values and their rankings. Therefore, we repeated this analysis five times and calculated the average SHAP values (feature importance scores). The top 20 most important features (Figures 2A and S12) include 16 chemical substructures, time, principle, guideline, and endpoint.

Reaction time was found to be the single most important parameter on biodegradation with the importance score much higher than those of the second and third features (Figure 2A). Principle, guideline, and endpoint were ranked the 3rd, 9th, and 15th, respectively. For the feature "Principle," CO_2 evolution, closed respirometer, and closed bottle tests were found to have similar SHAP values, much smaller than those of the DOC die away (Figure 2B), indicating higher percentages of biodegradation observed using the DOC die away methods. This is because microorganisms can sometimes utilize a portion of the organic carbon from the test substance for their own growth and reproduction, resulting in lower DOC levels in the solution and therefore higher calculated biodegradation percentages. This is also the reason that the DOC-based standard methods usually set a threshold of 70% (not 60% for other principles) to classify the ready biodegradability.[42,43] In addition, adsorption onto the inoculum or the inner walls of the reactor may occur for some test substances, which sometimes cannot be completely corrected by the control tests and therefore leads to falsely high results. This might also be the reason for the much wider range of SHAP values for this principle compared to others.

For the feature "Guideline" (Figure 2C), all ready biodegradation guidelines except for those based on DOC die away were observed to have similar SHAP values. This is reasonable because they were originally designed to be equivalent except that they can handle chemicals with different physicochemical properties, such as solubility, volatility, and absorptivity.[42,43] Although the rest five guidelines had noticeably higher SHAP values, all of them were either inherent biodegradation guidelines (i.e., OECD 302B and 302C) or the ready ones based on DOC die away (i.e., EU C.4-A, OECD 301A and 301E). For the feature "Endpoint"

https://doi.org/10.1021/acs.est.2c01764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

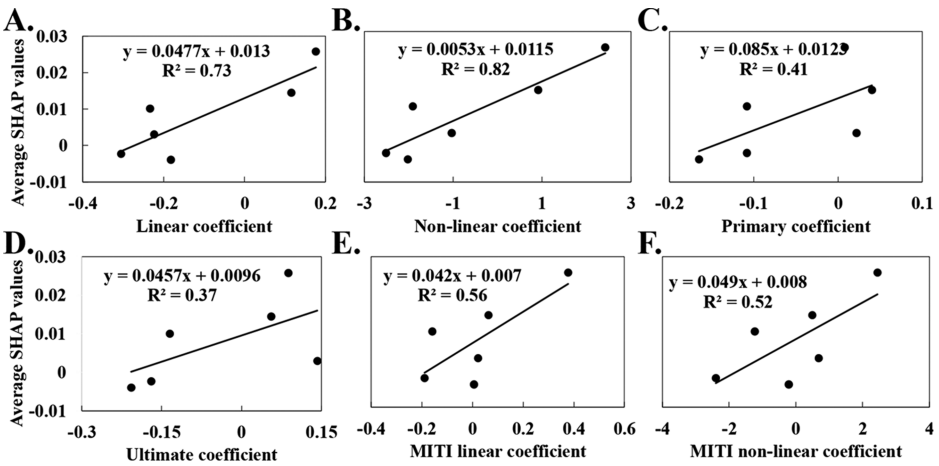


Figure 3. Plots of the average SHAP values of six common substituents on aromatic chemicals against six sets of group contribution coefficients for these substituents collected from the reported (A) linear, (B) nonlinear, (C) primary, (D) ultimate, (E) MITI linear, and (F) MITI nonlinear biodegradation models.[27,49]

(Figure 2D), the "Inherent" tests showed much higher SHAP values and therefore stronger effects on the observed biodegradation percentages than the "Ready" tests. This is expected because inherent tests usually have higher biomass to test substance ratios than those of the ready tests.[42,43] Please see more discussion later.

As shown in Figure 2A, all 16 chemical substructures that were ranked among the top 20 were unambiguously classified to have either positive or negative effects on the biodegradation. Such results (positive or negative) were indicated by the individual SHAP values shown in Figure S12. Among the 16 substructures, only three were found to have positive effects, that is, C═O (#154), OCO (#123), and OH (#139). This agrees well with the literature that the presence of carboxyl or hydroxyl groups can substantially improve the biodegradability.[22,44] The following substructures were found to have negative effects on biodegradation: rings (#163, 145, 165, 150), branches (#156, 112, 148, 66), halogens (#134), nitrogen substituents (#156, 142, 161), heteroatoms (#138, 148), and methyl group (#149). Most of them can lower the chemical hydrophilicity and/or electron density and have been widely reported to have negative effects on aerobic biodegradation.[22,44−47] From the most positive to the most negative, these substructures were ranked following the order of C═O > O═C−O > OH > CH_3 > halogen > branching > N > 6-member ring. Similarly, Boethling et al. reported an order of ester, amide, anhydride > hydroxyl > carboxyl, epoxide, site of unsaturation > benzene ring, methyl methylene.[45] Therefore, we believe that this model correctly identified the effects of substructures on aerobic biodegradation.

The group contribution method has been widely used to quantify the contributions of functional groups toward aerobic biodegradation.[27,48] To better evaluate the accuracy of the mechanisms learned by our model, we further compared the substructure SHAP values with the reported substructure coefficients in the group contribution method.[27] A set of common substituents on aromatic chemicals (i.e., COOH, OH, Cl, NO_2, NH_2 or NH, and SO_3H) and aliphatic chemicals (i.e., COOH, OH, Cl, NH_2 or NH, and SO_3H) were evaluated. As most of them appeared in a few hundred to over 3000 records in the dataset, the SHAP values for each substituent were averaged. A total of six sets of coefficients for these substituents were collected from the reported linear, nonlinear, primary, ultimate, MITI linear, and MITI nonlinear biodegradation models.[27,49] Note that these models were developed to predict the probability/time frame of biodegradation based on the group contributions of different substituents and are the basis of the US EPA's software EPI Suite for predicting biodegradability.[18] The linear regression of these averaged SHAP values against the six sets of coefficients had R^2 values of 0.73, 0.82, 0.41, 0.37, 0.56, and 0.52, respectively, for the substituents on aromatic chemicals (Figure 3) and 0.12, 0.05, 0.70, 0.65, 0.92, and 0.91, respectively, for the substituents on aliphatic chemicals (Figure S13). The detailed coefficient and the average SHAP values can be found in Tables S9 and S10. Given the much larger number of chemicals in this study than in the group contribution method (>6000 vs <300), these mostly decent correlations suggest that our model was likely based on a correct understanding of the substructure contributions/importance to aerobic biodegradation, and is, hence, trustworthy.

To have a better idea of the occurrence of MACCS substructures in our database, we summarized the number of records that have different MACCS bits, as shown in Figure S14 and Table S11. The predictions for the chemicals that have at least 10% of their total bits to be among the 20 most popular bits (Table S11) achieved an RMSE and an R^2 of 0.23 and 0.58, respectively. However, these values changed to 0.27 and 0.34, respectively, for the chemicals that have at least 10% of their total bits to be among the 20 least popular bits (Table S11). This indicates that the model performed better for chemicals containing substructures that occur more frequently. Future studies may consider adding chemicals containing substructures that have low occurrence frequencies in our current database so that their contributions/importance to aerobic biodegradation can be better understood. This can also in turn expand the model AD.

Individual Ready and Inherent Models, and Knowledge Transfer. The models built so far considered both ready and inherent biodegradation data. To examine whether building separate models based on the ready or inherent data alone could give better prediction accuracy for either endpoint, we split the original dataset into ready and inherent subsets and built a ready and an inherent model separately (details in Text S1.10). The performance of the obtained ready

https://doi.org/10.1021/acs.est.2c01764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

model was similar to that of the overall model for predicting ready biodegradation (data not shown). This might be because the original dataset on which the overall model was built consisted of mostly ready biodegradation data (90%, Table S2). As shown in Figure S15, the inherent model only showed marginal improvement (<0.01 for both RMSE and R^2 on the scale of 0.0−1.0, with the p values of 0.48 and 0.29 (>0.05), respectively, in ANOVA analyses) over the overall model for the inherent biodegradability.

When knowledge transfer (using the ready model to predict the ready biodegradation for the chemicals in the inherent dataset and then using the predicted results as an additional input feature in the inherent model) (details in Text S1.10), poorer model performance was observed compared to the overall model. Note that employing the data augmentation technique did not improve the model performance (Text S1.10 and Figure S15). Discussion on the possible reasons can be found in Text S2.3 and Figures S16 and S17.

As the individual ready and inherent models did not significantly outperform the overall model, the overall model was used for the rest of the study.

Classification Models. In addition to regression models, we developed classification models to expand the application of this study to when only binary classification is needed (an industry norm) without specifying the experimental conditions and reaction time. For the first classification model based on the dataset ClassDataset_original (data extracted from the regression dataset), data balancing was first performed (Figure S18). With MACCS FPs as the chemical representation, a total of 14 ML algorithms (Table S5) were examined, and XGBoost was again found to be one of the best algorithms (Figure 4). Given that it has been used for the regression models, we selected it as our default algorithm for the classification models.

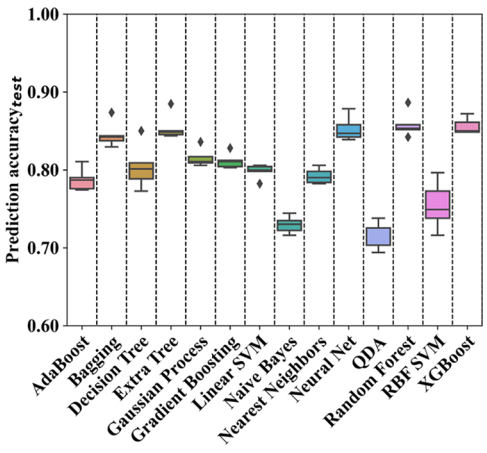


Figure 4. Comparison of the performance of 14 ML algorithms using their default hyperparameters for the development of classification models.

When Bayesian optimization was performed for the hyperparameter tuning, we did not observe improvement in the model performance. Therefore, the default values were used (Table S12). The model yielded prediction accuracy, sensitivity, specificity, balanced accuracy, AUC, and f_1 score of 85.6, 90.3, 80.4, 85.4, 91.8, and 85.1%, respectively, on the test set (Table S13), better than most of the reported models in the literature.[1,8,12,21,24] For example, a recent study with the most comprehensive datasets (3146 chemicals) ever only achieved balanced accuracies of 72−75%.[24] Many other studies with much smaller datasets showed sensitivities of 61−85% and specificities of 80−93%.[1,4,12,21]

To evaluate the model performance on the external dataset ClassDataset_external (data collected from the literature, containing 537 1s and 1009 0s), we calculated the similarities between the chemicals in this dataset and those in the ClassDataset_original (used for the model development), as shown in Figure S19. Compared to the chemicals in the ClassDataset_original, 96.9% of the chemicals in ClassDataset_external had similarity scores over 0.6, suggesting high reliability of the prediction. Indeed, the prediction accuracy of 84.5% was observed for ClassDataset_external, very close to the value for the test dataset of ClassDataset_original.

The above two datasets were then combined to form ClassDataset_all, which was used to build a more comprehensive classification model called ClassModel_all so that it can have a larger AD. Following the same procedure, a prediction accuracy was calculated to be 86.0% for the test set, similar to that of the ClassModel_original. The results on other matrices such as sensitivity and specificity can be found in Table S13. Overall, these results were very similar to those obtained for the model ClassModel_original. This model is used for the rest of the study unless otherwise noted.

Effects of Chemical Speciation on Biodegradability. To examine the effect of chemical acid/base speciation on the predicted biodegradability, we compared the model performance with two sets of pK_a values and/or α notations for each chemical added to the input features: (1) 20 pK_a values (10 for acids and 10 for bases) and 22 α notations and (2) 8 pK_a values (4 for acids and 4 for bases) and 10 α notations. Because adding either set of the features had similar effects on the model performance (results not shown), we did not try to decrease the number of pK_a values or α notations further. The results of adding 8 pK_a values and 10 α notations are discussed below.

As shown in Figure 5A, surprisingly, the inclusion of pK_a values and/or α notations did not noticeably improve the performance of the regression model (reasons unknown). However, for the classification model built on the dataset ClassDataset_all (Figure 5B), the performance was significantly improved, as indicated by most of the matrices used in this study, including accuracy (from 85.1 to 87.6%), specificity (from 80.9 to 87.4%), balanced accuracy (from 84.9 to 87.6%), AUC (from 92.4 to 94.8%), and f_1 (from 86.2 to 87.9%). The model performance generally follows the order of "with both pK_a and α" > "with α" > "with pK_a" > "without pK_a or α." This finding demonstrates the importance of chemical speciation in aerobic biodegradation kinetics.[50]

Model AD and Application to the Database DSSTox. The AD of the regression model was determined and shown in Table 1. The expected prediction accuracy of a query compound can be found based on its similarity to the model dataset. For example, if a chemical has a similarity score higher than 0.9, its expected prediction accuracy can be represented by an RMSE of 0.14 and an R^2 of 0.79. Chemicals with the similarity score lower than 0.5 was defined as out of AD because we did not have enough test chemicals in this range to confidently evaluate the model performance (details in Text S1.12). Similarly, the AD of the classification model can be found in Table S15.

https://doi.org/10.1021/acs.est.2c01764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

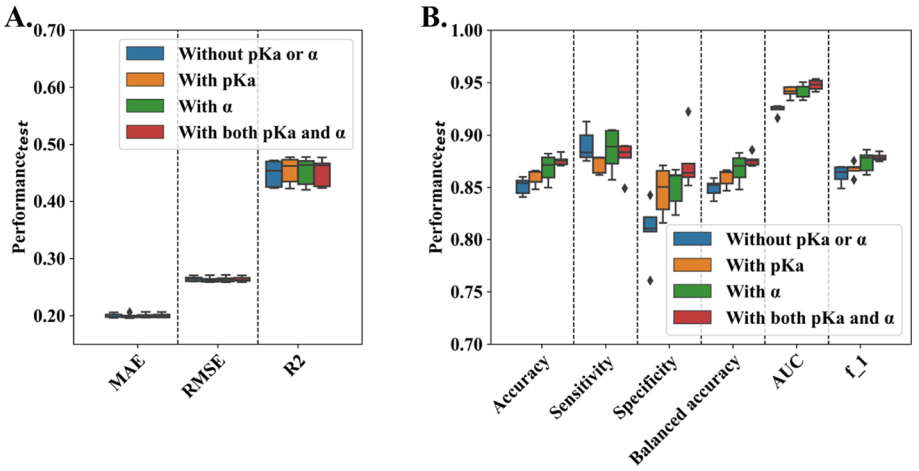


Figure 5. Model performance with or without pK_a and/or α notations of the chemicals for the (A) regression model (built on all the 12,750 data points) and (B) classification model (built on the dataset ClassDataset_all). Detailed values for (B) can be found in Table S14.

Table 1. AD of the Regression Model and the Model Applicability Toward DSSTox

	Expected prediction		Chemical percentages in DSSTox	
Similarity	RMSE	R^2	Each level (%)	Accumulative (%)
0.9−1.0	0.14	0.79	7.1	7.1
0.8−0.9	0.21	0.66	15.4	22.5
0.7−0.8	0.23	0.59	31.5	54.0
0.6−0.7	0.26	0.44	34.1	88.1
0.5−0.6	**0.26**	**0.49**	**10.4**	**98.5**
0.4−0.5	Out of AD	Out of AD	1.0	99.5
<0.4	Out of AD	Out of AD	0.5	100.0

As shown in Figure S20 and Table 1, for the regression model, the similarity calculation shows that 98.5% of the chemicals in the DSSTox fall into the model AD; 6.7% and 93.3% were predicted to have ≥60% and <60% of biodegradation, respectively, under the guideline of OECD 301F over 28 days (other model input: "ready" as the endpoint, "closed respirometer" as the Principle, and "1" as the reliability). For the classification model, 98.4% of the chemicals in the DSSTox are within the model AD, and 88.2% were classified with an expected accuracy of 85.6−88.9% (Figure S21 and Table S15). 15.6% and 84.4% were classified to be RB and NRB, respectively.

When the prediction results of both the regression and classification models were compared, 6.1% of the same DSSTox chemicals were found to have either ≥60% of biodegradation (by the regression model) or a "1" (RB) (by the classification model), while 83.8% of the same DSSTox chemicals were predicted to have either <60% of biodegradation or a "0" (NRB). These results demonstrate that the two models had consistent prediction results for 89.9% of the DSSTox chemicals. The similarity calculation further revealed that the 10.1% inconsistent predictions were mostly for chemicals having low similarity scores (i.e., a larger percentage of chemicals out of AD) (Figure S22). The detailed prediction results for each chemical can be found in the Excel file "DSSTox prediction.xlsx" in the Supporting Information.

Free Online Predictors. To make the models developed in this study more accessible, we developed two free online predictors for the regression and classification models (https://www.chemai.aropha.com/). Although pK_a values can be predicted, they may require further treatment depending on the nature of the chemicals. This made it difficult to include pK_a in these two predictors. Therefore, the regression and classification models obtained right before considering pK_a and α notations were used to develop our online predictors. The predictors accept direct SMILES strings or Excel/CSV files containing SMILES strings in the column named "SMILES" or SDF files as the input. For the regression predictor, additional inputs such as the guideline and principle should also be provided (select from the dropdown options). Upon submit, the predictors make predictions and at the same time calculate the similarities between the query chemicals and the datasets of the models to evaluate the prediction accuracies. All results are shown in a downloadable table. In addition, users can download all related datasets used in this study and the resulted model files on the website. With the model files, users can follow our Jupyter Notebooks step by step to perform predictions using Python.

■ ENVIRONMENTAL IMPLICATIONS

This study developed regression and classification models using ML for the prediction of aerobic biodegradability of organic chemicals in water. A total of 12,750 data points were collected for the regression model, which is substantially larger than others reported in the literature (always less than 3200). This significantly improved the model AD. Different from most other studies where only classification models were evaluated, the developed regression models considerably improved the prediction accuracy by changing the output from 0 (NRB) and 1 (RB) to continuous biodegradation percentages (0−100%). This also helped cover more structure−biodegradability relationships. The inclusion of guidelines, principles, endpoints, and time as additional inputs for the first time significantly improved the model applicability by providing practical options for users to, for example, compare the prediction results under different guidelines or principles or obtain the full biodegradation kinetics from days 0 to 28 (or up to 73 d). Also, for the first time, the prediction of inherent biodegradability is available by including more than 1270 inherent data points in the model. This is especially helpful for chemicals that are known/predicted to be NRB as they can now be predicted under inherent biodegradation conditions without having to be tested experimentally.

https://doi.org/10.1021/acs.est.2c01764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

The classification model developed based on more than 6000 chemicals showed high robustness and higher prediction accuracies than most of the reported models. Both the regression and classification models have the largest ADs ever reported, large enough to cover more than 98% of the 850,000 environmentally relevant chemicals in the database DSSTox. The two freely available online predictors made these models readily useable even for users who have little ML knowledge. With the significantly increased prediction accuracies and enlarged ADs compared to others, these models can also help provide more accurate risk assessment of the existing and new chemicals. We believe that the model predicted biodegradability for the chemicals in the DSSTox database and the free online predictors can effectively help the research community, chemical industries, and regulators more easily achieve their research, production, and regulatory goals.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.2c01764.

Datasets used in this study (XLSX)

DSSTox dataset prediction (XLSX)

More details about dataset preparation, chemical representation, other inputs and categorical data encoding, ML algorithms, model development, feature importance evaluation by SHAP analysis, knowledge transfer, effect of speciation, model AD, and supplementary figures and tables (PDF)

Model files and sample codes in Jupyter Notebooks (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Huichun Zhang − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States;* orcid.org/0000-0002-5683-5117; Phone: (216) 368-0689; Email: hjz13@case.edu

Author

Kuan Huang − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States;* orcid.org/0000-0003-4657-4686

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.2c01764

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the National Science Foundation grant CHE-2105005. H.Z. acknowledges the Lubrizol Corporation for helpful discussions and their support in the initial phase of this work.

■ REFERENCES

(1) Cheng, F.; Ikenaga, Y.; Zhou, Y.; Yu, Y.; Li, W.; Shen, J.; Du, Z.; Chen, L.; Xu, C.; Liu, G.; Lee, P. W.; Tang, Y. In silico assessment of chemical biodegradability. *J. Chem. Inf. Model* **2012**, *52*, 655−669.

(2) Eubeler, J. P.; Zok, S.; Bernhard, M.; Knepper, T. P. Environmental biodegradation of synthetic polymers I. Test methodologies and procedures. *TrAC, Trends Anal. Chem.* **2009**, *28*, 1057−1072.

(3) Aropha. Biodegradation Test Method Overview: Aropha Resource Center; Aropha Inc., 2022. https://www.resources.aropha.com/docs/test-methods/biodegradation-test-method-overview/ (accessed 06/20/2022).

(4) Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative structure-activity relationship models for ready biodegradability of chemicals. *J. Chem. Inf. Model* **2013**, *53*, 867−878.

(5) Fernández, A.; Rallo, R.; Giralt, F. Prioritization of in silico models and molecular descriptors for the assessment of ready biodegradability. *Environ. Res.* **2015**, *142*, 161−168.

(6) Pizzo, F.; Lombardo, A.; Brandt, M.; Manganaro, A.; Benfenati, E. A new integrated in silico strategy for the assessment and prioritization of persistence of chemicals under REACH. *Environ. Int.* **2016**, *88*, 250−260.

(7) Howard, P. H.; Boethling, R. S.; Stiteler, W.; Meylan, W.; Beauman, J. Development of a predictive model for biodegradability based on BIODEG, the evaluated biodegradation data base. *Sci. Total Environ.* **1991**, *109−110*, 635−641.

(8) Tunkel, J.; Howard, P. H.; Boethling, R. S.; Stiteler, W.; Loonen, H. Predicting ready biodegradability in the Japanese Ministry of International Trade and Industry test. *Environ. Toxicol. Chem.* **2000**, *19*, 2478−2485.

(9) Hiromatsu, K.; Yakabe, Y.; Katagiri, K.; Nishihara, T. Prediction for biodegradability of chemicals by an empirical flowchart. *Chemosphere* **2000**, *41*, 1749−1754.

(10) Cuissart, B.; Touffet, F.; Crémilleux, B.; Bureau, R.; Rault, S. The maximum common substructure as a molecular depiction in a supervised classification context: Experiments in quantitative structure/biodegradability relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1043−1052.

(11) Philipp, B.; Hoff, M.; Germa, F.; Schink, B.; Beimborn, D.; Mersch-Sundermann, V. Biochemical Interpretation of Quantitative Structure−Activity Relationships (QSAR) for Biodegradation of N-Heterocycles: A Complementary Approach to Predict Biodegradability. *Environ. Sci. Technol.* **2007**, *41*, 1390−1398.

(12) Toropov, A.; Toropova, A.; Lombardo, A.; Roncaglioni, A.; Brita, N.; Stella, G.; Benfenati, E. CORAL: the prediction of biodegradation of organic compounds with optimal SMILES-based descriptors. *Open Chem.* **2012**, *10*, 1042.

(13) Pizzo, F.; Lombardo, A.; Manganaro, A.; Benfenati, E. In silico models for predicting ready biodegradability under REACH: a comparative study. *Sci. Total Environ.* **2013**, *463−464*, 161−168.

(14) Chen, G.; Li, X.; Chen, J.; Zhang, Y. N.; Peijnenburg, W. J. Comparative study of biodegradability prediction of chemicals using decision trees, functional trees, and logistic regression. *Environ. Toxicol. Chem.* **2014**, *33*, 2688−2693.

(15) Lombardo, A.; Pizzo, F.; Benfenati, E.; Manganaro, A.; Ferrari, T.; Gini, G. A new in silico classification model for ready biodegradability, based on molecular fragments. *Chemosphere* **2014**, *108*, 10−16.

(16) Vorberg, S.; Tetko, I. V. Modeling the biodegradability of chemical compounds using the Online CHEmical Modeling Environment (OCHEM). *Mol. Inf.* **2014**, *33*, 73−85.

(17) Ballabio, D.; Biganzoli, F.; Todeschini, R.; Consonni, V. Qualitative consensus of QSAR ready biodegradability predictions. *Toxicol. Environ. Chem.* **2017**, *99*, 1193−1216.

(18) Zhan, Z.; Li, L.; Tian, S.; Zhen, X.; Li, Y. Prediction of chemical biodegradability using computational methods. *Mol. Simul.* **2017**, *43*, 1277−1290.

(19) Toropov, A. A.; Toropova, A. P. Improved model for biodegradability of organic compounds: the correlation contributions of rings. *Toxicity and Biodegradation Testing*; Springer, 2018; pp 147−183.

(20) Nolte, T. M.; Pinto-Gil, K.; Hendriks, A. J.; Ragas, A. M. J.; Pastor, M. Quantitative structure-activity relationships for primary aerobic biodegradation of organic chemicals in pristine surface waters:

starting points for predicting biodegradation under acclimatization. *Environ. Sci.: Processes Impacts* **2018**, *20*, 157−170.
(21) Putra, R. I. D.; Maulana, A. L.; Saputro, A. G. Study on building machine learning model to predict biodegradable-ready materials. *AIP Conf. Proc.* **2019**, *2088*, 060003.
(22) Acharya, K.; Werner, D.; Dolfing, J.; Barycki, M.; Meynet, P.; Mrozik, W.; Komolafe, O.; Puzyn, T.; Davenport, R. J. A quantitative structure-biodegradation relationship (QSBR) approach to predict biodegradation rates of aromatic chemicals. *Water Res.* **2019**, *157*, 181−190.
(23) Tang, W.; Li, Y.; Yu, Y.; Wang, Z.; Xu, T.; Chen, J.; Lin, J.; Li, X. Development of models predicting biodegradation rate rating with multiple linear regression and support vector machine algorithms. *Chemosphere* **2020**, *253*, 126666.
(24) Lunghini, F.; Marcou, G.; Gantzer, P.; Azam, P.; Horvath, D.; Van Miert, E.; Varnek, A. Modelling of ready biodegradability based on combined public and industrial data sources. *SAR QSAR Environ. Res.* **2020**, *31*, 171−186.
(25) Niemi, G. J.; Veith, G. D.; Regal, R. R.; Vaishnav, D. D. Structural features associated with degradable and persistent chemicals. *Environ. Toxicol. Chem.* **1987**, *6*, 515−527.
(26) Boethling, R. S. Application of molecular topology to quantitative structure-biodegradability relationships. *Environ. Toxicol. Chem.* **1986**, *5*, 797−806.
(27) Boethling, R. S.; Howard, P. H.; Meylan, W.; Stiteler, W.; Beauman, J.; Tirado, N. Group contribution method for predicting probability and rate of aerobic biodegradation. *Environ. Sci. Technol.* **1994**, *28*, 459−465.
(28) Jaworska, J. S.; Boethling, R. S.; Howard, P. H. Recent developments in broadly applicable structure-biodegradability relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 1710−1723.
(29) Raymond, J. W.; Rogers, T. N.; Shonnard, D. R.; Kline, A. A. A review of structure-based biodegradation estimation methods. *J. Hazard. Mater.* **2001**, *84*, 189−215.
(30) Rücker, C.; Kümmerer, K. Modeling and predicting aquatic aerobic biodegradation−a review from a user's perspective. *Green Chem.* **2012**, *14*, 875−887.
(31) Sabljic, A.; Nakagawa, Y., Biodegradation and Quantitative Structure-Activity Relationship (QSAR). *Non-First Order Degradation and Time-dependent Sorption of Organic Chemicals in Soil*; American Chemical Society, 2014; Vol. *1174*, pp 57−84.
(32) Singh, A. K.; Bilal, M.; Iqbal, H. M. N.; Raj, A. Trends in predictive biodegradation for sustainable mitigation of environmental pollutants: Recent progress and future outlook. *Sci. Total Environ.* **2021**, *770*, 144561.
(33) 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.
(34) Zhong, S.; Hu, J.; Fan, X.; Yu, X.; Zhang, H. A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants. *J. Hazard. Mater.* **2020**, *383*, 121141.
(35) Yang, H.; Huang, K.; Zhang, K.; Weng, Q.; Zhang, H.; Wang, F. Predicting Heavy Metal Adsorption on Soil with Machine Learning and Mapping Global Distribution of Soil Adsorption Capacities. *Environ. Sci. Technol.* **2021**, *55*, 14316−14328.
(36) Zhong, S.; Zhang, K.; Wang, D.; Zhang, H. Shedding Light On "Black Box" Machine Learning Models for Predicting the Reactivity of HO• Radicals toward Organic Compounds. *Chem. Eng. J.* **2020**, *405*, 126627.
(37) Zhong, S.; Zhang, Y.; Zhang, H. Machine Learning-Assisted QSAR Models on Contaminant Reactivity Toward Four Oxidants: Combining Small Data Sets and Knowledge Transfer. *Environ. Sci. Technol.* **2022**, *56*, 681−692.
(38) Pan, X.; Wang, H.; Li, C.; Zhang, J. Z. H.; Ji, C. MolGpka: A Web Server for Small Molecule pKa Prediction Using a Graph-Convolutional Neural Network. *J. Chem. Inf. Model.* **2021**, *61*, 3159−3165.
(39) Martin, T. J.; Snape, J. R.; Bartram, A.; Robson, A.; Acharya, K.; Davenport, R. J. Environmentally relevant inoculum concentrations improve the reliability of persistent assessments in biodegradation screening tests. *Environ. Sci. Technol.* **2017**, *51*, 3065−3073.
(40) Cvetnic, M.; Juretic Perisic, D.; Kovacic, M.; Kusic, H.; Dermadi, J.; Horvat, S.; Bolanca, T.; Marin, V.; Karamanis, P.; Bozic, A. L. Prediction of biodegradability of aromatics in water using QSAR modeling. *Ecotoxicol. Environ. Saf.* **2017**, *139*, 139−149.
(41) 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.
(42) OECD. *OECD Guideline for Testing of Chemicals, Test No. 301: Ready Biodegradability*; OECD, 1992.
(43) OECD. *Test No. 302B: Inherent Biodegradability*; Zahn-Wellens/EVPA Test, 1992.
(44) Acharya, K.; Werner, D.; Dolfing, J.; Meynet, P.; Tabraiz, S.; Baluja, M. Q.; Petropoulos, E.; Mrozik, W.; Davenport, R. J. The experimental determination of reliable biodegradation rates for mono-aromatics towards evaluating QSBR models. *Water Res.* **2019**, *160*, 278−287.
(45) Boethling, R. S.; Gregg, B.; Frederick, R.; Gabel, N. W.; Campbell, S. E.; Sabljic, A. Expert systems survey on biodegradation of xenobiotic chemicals. *Ecotoxicol. Environ. Saf.* **1989**, *18*, 252−267.
(46) Martin, T. J.; Goodhead, A. K.; Acharya, K.; Head, I. M.; Snape, J. R.; Davenport, R. J. High throughput biodegradation-screening test to prioritize and evaluate chemical biodegradability. *Environ. Sci. Technol.* **2017**, *51*, 7236−7244.
(47) Pitter, P.; Chudoba, J. *Biodegradability of Organic Substance in the Aquatic Environment*; AGRIS, 1990.
(48) Jhamb, S.; Hospital, I.; Liang, X.; Pilloud, F.; Piccione, P. M.; Kontogeorgis, G. M. Group Contribution Method to Estimate the Biodegradability of Organic Compounds. *Ind. Eng. Chem. Res.* **2020**, *59*, 20916−20928.
(49) Tunkel, J.; Howard, P. H.; Boethling, R. S.; Stiteler, W.; Loonen, H. Predicting ready biodegradability in the Japanese Ministry of International Trade and Industry test. *Environ. Toxicol. Chem. Int. J.* **2000**, *19*, 2478−2485.
(50) Nolte, T. M.; Pinto-Gil, K.; Hendriks, A. J.; Ragas, A. M.; Pastor, M. Quantitative structure−activity relationships for primary aerobic biodegradation of organic chemicals in pristine surface waters: starting points for predicting biodegradation under acclimatization. *Environ. Sci.: Processes Impacts* **2018**, *20*, 157−170.

https://doi.org/10.1021/acs.est.2c01764
Environ. Sci. Technol. XXXX, XXX, XXX−XXX