ACS ES&T | Water

pubs.acs.org/estwater Review

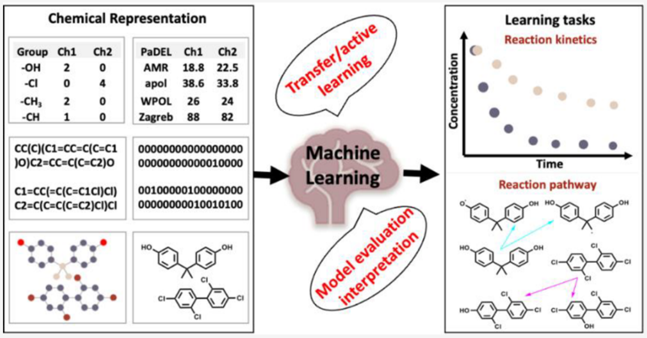
Machine Learning Modeling of Environmentally Relevant Chemical Reactions for Organic Compounds

Kai Zhang and Huichun Zhang*

Cite This: https://doi.org/10.1021/acsestwater.2c00193 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Environmental chemical reactions have been frequently investigated for various purposes; however, it remains challenging to accurately model either the reaction kinetics or reaction pathways. Existing studies mostly model reaction kinetics with traditional quantitative structure−activity relationships (QSARs) or reaction pathways with reaction template methods; however, these approaches generally require extensive feature engineering or manual extraction of reaction templates. Recently, machine learning (ML) has become a promising tool for modeling chemical reactions as ML models can perform well and are powerful in using diverse chemical representations. This Review starts with a concise comparison of traditional and ML modeling approaches for chemical reactions, followed by a brief discussion of the status of and future needs in modeling environmental organic reactions. Data collection and data cleaning techniques for reaction kinetics and pathways are then discussed. We then summarize the advantages and limitations of commonly used chemical representations and feature selection techniques. Next, we critically review general ML model evaluation and interpretation processes and propose a three-step evaluation process, that is, comparisons with general metrics, baseline models, and existing models. Lastly, we explore ML modeling approaches for small data sets, including transfer learning and active learning, which have been successfully employed in many other fields, for future modeling of environmental chemical reactions.



1. INTRODUCTION

Chemical reactions of organic contaminants, such as advanced oxidation processes (AOPs),[1] redox reactions,[2] photolysis,[3] photocatalytic reactions,[4] hydrolysis,[5] and biodegradation,[6] have been widely observed in environmental transformation and water treatment processes.[7−10] To understand these environmentally relevant chemical reactions, we need to consider two important aspects: the reaction kinetics (rates or rate constants) and the reaction products or pathways. The reaction kinetics are commonly obtained by measuring changes in the contaminant concentration over time; the reaction pathways are often understood after identifying the major reaction products and intermediates as well as establishing linkages among different species on the basis of known reaction rules. To examine either aspect, we have to rely on many experiments or computational calculations.

Due to a large number of emerging organic pollutants, various studies have attempted to build models to predict their reaction rate constants or products. Traditional modeling approaches mostly rely on quantitative structure−activity relationships (QSARs) or known reaction rules to predict rate constants or products for structurally similar compounds under similar conditions.[11] For example, the group contribution method[12−14] or the molecular descriptor has been widely used to build predictive models for the rate constants of hydroxyl radicals with organic contaminants;[15,16] reaction rules or templates extracted from the literature[8,17] have been employed to predict reaction pathways/products. However, kinetic modeling is often limited to linear correlations and requires extensive feature engineering—selecting the most relevant variables (features) as the model input.[11] Template-based product predictions heavily rely on handcrafted reaction rules, which are often of questionable quality and can be hard to scale up to handle new reactions and compounds.[18] In addition, these approaches cannot well consider the effects of reaction conditions like temperature and pH,[12−16] which are known to strongly affect the kinetics and sometimes reaction products,[19−21] so the corresponding models would inevitably fail to perform well when extended to different reaction

Special Issue: Applications of Artificial Intelligence, Machine Learning, and Data Analytics in Water Environments

Received: May 2, 2022
Revised: July 5, 2022
Accepted: July 6, 2022

Table 1. Summary of ML Models on Environmentally Relevant Chemical Reactions

modeling type	learning task	input/algorithm	performance	ref.
reaction rates	defluorination of per- and polyfluoroalkyl substances	molecule descriptors (MDs)/gradient boosting regression (GBR)	R^2 = 0.944; RMSE = 0.114	27
	defluorination energy of per- and polyfluoroalkyl compounds	chemical bond descriptors/RF, LASSO, and neural network (NN)	R^2 = 0.934; RMSE = 1.22	37
	organic pollutants with HClO, O_3, ClO_2, and $SO_4\cdot$	molecular fingerprints (MFs) and MD/NN	RMSE = 2.04 (HClO); 1.94 (O_3); 1.49 (ClO_2); 0.7 ($SO_4\cdot$)	21
	organic pollutants with ·OH and $SO_4\cdot$	MF/RF and NN	R^2 = 0.931/0.916 (·OH/$SO_4\cdot$); RMSE = 0.639−0.823/0.767−0.824	28
	organic pollutants with an Fe(II) complex	MF/RF	RMSE = 0.43	34
	organic pollutants with ·OH	images/convolutional neural network (CNN)	RMSE = 0.123−0.151/0.284−0.339 (train/test)	26
	organic pollutants with O_3	MD/support vector machine (SVM)	R^2 = 0.862/0.782 (train/test)	38
	organic pollutants with ·OH	MF/NN	R^2 = 0.972/0.789 (train/test) RMSE = 0.135/0.329	31
	organic pollutants with ·OH	MD and quantum descriptors/NN	R^2 = 0.848/0.879 (train/test) RMSE = 0.254/0.356	39
reaction pathways/ products	organic chemical reaction products	graph/graphic neural network (GNN)	accuracy = 85.6%	32
	reaction product prediction	SMILES/Transformer	accuracy = 90%	18
	organic reaction classification	MF/NN	accuracy = 86%/ 85.7% (train/test)	40

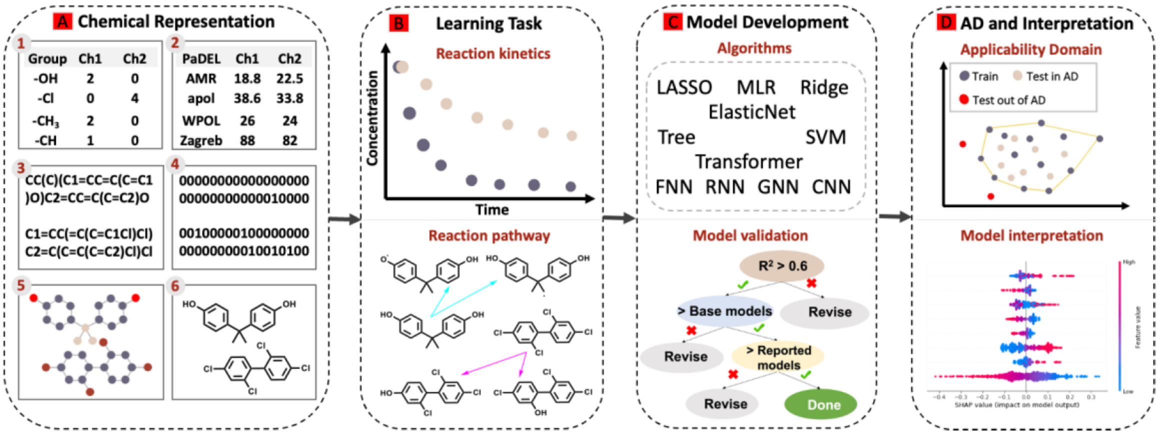
conditions. Also, the traditional QSAR approach mainly relies on manual feature selection and likely requires feature selection to be repeated whenever new data records are added. Such a recursive process makes it hard for later studies to conveniently build consistent models because different input features may be selected each time. In addition to QSARs, quantum chemical calculations have also been used to perform kinetic modeling;[22] however, these calculations generally require the enumeration of all possible reaction pathways and, hence, are challenging for complex environmentally relevant reactions.[23]

Recently, chemical reaction modeling using machine learning (ML) has gained much attention. Armed with more data and advanced algorithms, ML models have significantly outperformed traditional models and greatly simplified the modeling process.[24−28] One important requirement for building successful chemical models is to use proper representations of different compounds, where ML is particularly attractive as it can directly utilize a diverse range of chemical representations, such as molecular descriptors,[29] molecular fingerprints,[30,31] images,[26] strings,[18] and graphs,[32] which considerably reduces the requirements for feature engineering.[33] For example, recent studies have successfully predicted the reaction rate constants of hundreds of organic compounds toward either common oxidants, including hydroxyl radicals, sulfate radicals, HClO, ClO_2, and ozone, or Fe(II)-based reductants using molecular fingerprints and different ML algorithms[25,31,34] or using chemical images of molecules and a convolutional neural network (CNN).[26] For reaction pathway modeling, ML has also helped address long-lasting challenges such as exhaustive characterization of the state information in the reactive molecular collision and simulation of the complex reaction networks in combustion processes.[23] Also, ML can allow new modeling strategies such as building unified models or multitask models, which may achieve better prediction performance by building one "big" model covering several reaction systems rather than one model per system.[21,34] ML can also build and improve models in a more sustainable way. For example, when building models using chemical images or graphs,[26] later studies can upgrade existing models with new data by simply adding new chemical images or graphs into the training data set and retraining the models using the same training strategies and algorithms.

Despite these merits, existing models based on ML algorithms have some limitations compared with traditional ones, such as being more difficult to interpret and having overfitting risks on small data sets. For example, neural network models for rate constant predictions generally have multiple layers with many neurons in each layer.[25,35] As multiple matrix operations are performed on the input in each layer, it is challenging to directly analyze the trained models. In comparison, traditional models often use much simpler algorithms such as multiple linear regression (MLR) and can be conveniently interpreted. Besides, most existing studies on environmentally related reaction modeling focus on reaction kinetics, but only several studies have tried to model reaction pathways (Table 1). For a given environmental chemical transformation, the kinetics are certainly one of the most important considerations, especially for pollutant elimination purposes; however, the intermediates and final products would be another priority because the transformation of certain pollutants may produce more harmful products.[36]

In comparison, promising results have been reported for reaction modeling in organic chemistry, especially in modeling reaction pathways in organic synthesis and retrosynthesis. For example, studies have predicted reaction products using forward neural networks (FNNs) together with reaction templates,[40] graph neural networks with molecular graphs,[32] or the "Transformer" model.[18] Note that a Transformer treats reaction prediction as a translation process from reactants to products.[41] All these results have shown the promising potential of applying ML to model reaction products. Also, ML modeling in organic chemistry and environmental reaction studies faces similar issues. For example, the Transformer had a large number of parameters (tens of millions) and required a massive amount of data to ensure valid model training and validation,[18] while many reaction databases have very limited data records,[42] often less than the most studied ·OH-involved reactions (<2000 records).[21,26] To address data scarcity, ML studies on organic reactions have developed sophisticated strategies, such as transfer learning based on large general reaction databases, selectively enlarging databases through

Scheme 1. Flowchart of Building ML Models for Chemical Reactions[a]



[a](A) Selecting chemical representations from (1) group counting, (2) PaDEL descriptors, (3) SMILES, (4) molecular fingerprints, (5) molecular graphs, and (6) molecular images for two example compounds: Bisphenol A (Ch1) and PCB-47 (Ch2). (B) The two most common ML modeling tasks for chemical reactions: reaction kinetics (top) and reaction pathways (bottom). (C) Top: commonly used algorithms, ranging from simple linear regressions (MLR, LASSO, Ridge, and ElasticNet) to advanced decision tree/support vector machines (Tree and SVM) and to more sophisticated neural network algorithms (FNN, CNN, RNN, GNN, and Transformer). Bottom: the proposed three-step model evaluation in which the performance of an ML model is compared to a general metric such as R^2, then baseline models, and lastly reported models (Note that it is always helpful to conduct new experiments or calculations to further evaluate the model). The obtained ML model may need to be revised when it underperforms on the basis of any of the three comparisons. (D) Two additional important tasks after model development: defining the applicability domain (AD; top) and interpreting the model (bottom).

active learning, and taking advantage of unlabeled reactions.[42] Most of these techniques have yet to be employed in environmental reaction modeling. Given the similar challenges faced by both environmental reactions and organic reactions, it is very likely that environmental reaction modeling would greatly benefit from these techniques. Nevertheless, we have noticed that environmental reactions are often much more complex and may include multiple elementary reactions, with byproducts in each step. These complex reactions may be much harder to collect and build ML models for than some single-step organic synthesis reactions.

As there are recent reviews about ML algorithms[43] and general practices for building ML models,[25] this Review focuses on chemical reaction modeling. Specifically, we discuss the related challenges and future needs and provide possible solutions to these challenges/needs. In addition to model development itself, adequate data collection to build large, representative data sets is an important part of building robust, widely applicable predictive models. To this end, a few new studies and practices that are helpful for collecting reaction-related data are also discussed in this Review. Overall, the topics covered in this Review include data collection and data cleaning (Section 2); chemical representations, feature selection, and learning tasks (Section 3); model development, evaluation, and model interpretation (Section 4); and techniques for modeling small data sets (Section 5).

2. DATA COLLECTION AND DATA CLEANING

Literature data for reaction kinetics can either be directly collected by selecting reported rate constant values or be derived after fitting reaction kinetic curves from the literature, during which proper quality control is needed to ensure the data quality by, for example, setting well-defined criteria such as high R^2 values of the kinetic data, having reasonable control experiments, etc.[31] However, reaction pathways are much harder to collect because they are commonly reported in the form of images or complex chemical image networks, which include images of reactants, images of products, condition parameters, and arrows showing the directions of the reactions. Currently, most reaction pathway data sets in the environmental field rely on manual curation, which requires researchers to have expert knowledge about various reactions, is labor intensive, and is prone to mistakes. Researchers have to take a snapshot of each chemical in a reaction, draw their chemical structures, convert the structures to the SMILES, and assign these chemicals as reactants, products, or solvents according to the reaction network. This complex data collection process makes it hard to scale up to a large number of reactions. The rapid development of ML in computer vision has provided some very useful tools. For example, one study built a convolutional neural network model to automatically convert chemical images to SMILES, so one can conveniently obtain the SMILES of all involved chemicals by uploading a snapshot of the desired reaction.[44] With more such tools being developed, reaction data collection can be greatly facilitated. In addition, crowdsourcing, which has been widely used for data labeling or collection in computer vision and natural language processing, can be another way to collect reaction data. Future studies may take advantage of such an interesting tool; however, existing crowdsourcing platforms mostly focus on simple labeling tasks and may lead to significantly higher costs if one aims to label and collect complex chemical reactions.

Another important process during data preparation is cleaning of the raw data. The major reason for data cleaning and preprocessing is that most raw data sets include some redundant or even erroneous information, such as multiple records representing the same reaction or wrong SMILES for reactants or products in the collected raw data, so these records cannot contribute to the final model or even pose negative impacts on the model performance. A manual inspection of the data records is generally acceptable when the data set is not too large but is inefficient or even impossible when the sample size is large. For redundant information, it is preferable to canonicalize the SMILES (converting different

SMILES of one molecule from various sources into one single unique SMILES) of both reactants and products or use some unique identifier (e.g., CID from PubChem) for the involved chemicals because existing reactions may be misclassified as new ones if different SMILES from different sources are used for the same chemicals. To address these challenges, studies have proposed ML models to detect wrong information for organic reactions. For example, one recent study has built a Transformer model to clean the chemical reaction database and achieved well-improved results after removing incorrect reaction records.[45] Besides, commonly used chemical packages like Rdkit have built-in reaction-related modules (e.g., https://www.rdkit.org/docs/source/rdkit.Chem.rdChemReactions.html).[30] One can use such a module to validate the SMILES of the reactants and products within one reaction as a prescreening measure.

3. CHEMICAL REPRESENTATIONS, FEATURE SELECTION, AND LEARNING TASKS

Commonly used molecular descriptors (Scheme 1A) are calculated by packages like Dragon,[46,47] PaDEL,[29,48] RDKit,[30,34] and semiquantum MOPAC[49,50] and have been widely used in both traditional and ML modeling as the chemical representation. However, ML algorithms have greatly expanded the options of chemical representations to molecular fingerprints,[30,31] molecular graphs,[32] strings,[18] and even images.[33] Among these new chemical representations, molecular fingerprints contain 1s and 0s to indicate the presence or absence of certain functional groups or structures; molecular graphs use edges and nodes to represent atoms and bonds in molecules, while images use pixels to represent molecules (e.g., short lines to represent bonds and characters/colored balls to represent atoms).[30] Such versatile chemical representations have greatly reduced the requirement for input feature engineering, as one does not need to provide explicit reasoning for which features to select. This strengthens the ability of ML to model the underlying relationships beyond simple linear relationships. Besides, ML provides new opportunities to further improve the model performance, such as data augmentation. For example, molecular images used in CNN can be augmented by flipping or rotating images; the SMILES for "Transformer" can be augmented by randomization (using different SMILESs to represent one compound). Both methods have made ML models more robust for the test compounds.[30,51]

Despite these advantages, each chemical representation has its limitations. The molecular fingerprint is straightforward in showing specific structures but may lack 2D and 3D chemical information. Molecular descriptors include some 2D/3D features but may not always be available due to copyrights or missing values for certain compounds. Also, some packages are not open-sourced, so how certain descriptors are calculated is unknown. A molecular graph can easily describe atoms or bonds in a molecule but cannot effectively indicate the bond length or some 3D structural features of the compound, which would inevitably lead to the loss of chemical information.[52] Similarly, 2D images could not deliver 3D chemical information, and models based on images can be much harder to interpret than those using numbers or molecular graphs.[53] In addition to the specific limitations of each representation, all the above chemical representations suffer a similar problem; that is, they do not directly provide learning task-specific chemical information, such as the reaction mechanisms. Moreover, the more abstract the chemical representation, such as chemical images, the more complex is the algorithm to successfully train ML models, and more data will be required accordingly. Overall, there is no best chemical representation for all modeling tasks, and it would be better to evaluate the performance of different chemical representations on the same training/validation data set before the optimal one is selected.[54] In addition to difficulties in choosing representations for major chemicals in reactions, another challenge is to find suitable representations for substances that are involved in the reaction but cannot be easily quantified or described, such as radicals in advanced oxidation processes. One way to solve this is to use a categorical feature to represent different radicals.[21]

After the chemical representation has been selected, it may be helpful to reduce the number of input features, especially when a large number of molecular descriptors are used. Although ML algorithms are good at extracting useful information from high-dimensional inputs, a simpler input is still attractive because it can not only increase the efficiency of model training but also simplify the interpretation and application of the models. Commonly used feature selection approaches include using domain knowledge, using correlation coefficients either between input features and outputs or among input features,[55] conducting principal component analysis (PCA),[56,57] etc. The autoencoder,[58] which includes one encoder and one decoder with the output of the decoder being the same as the input of the encoder, can also help reduce the number of input features but still keep the most essential chemical information. For example, one research study used an autoencoder with the "SMILES" as the input to model chemical properties for the ZINC and PM9 data sets and achieved satisfactory results.[59]

Currently, chemical reaction modeling focuses on two major learning tasks, reaction kinetics and reaction pathways (Scheme 1B). A recent trend in modeling reaction kinetics is that more and more reactants are modeled. For example, early ML modeling may only focus on one reactive species like OH· or ozone,[38,39] while recent studies expand the modeling work to additional reactive species such as HClO, ClO_2, and SO_4·.[21,28] Meanwhile, earlier studies paid little attention to reaction conditions such as temperature and pH and treating compounds as the only variables; however, recent studies have tried to include these conditional parameters into the models.[21] As for the reaction pathway modeling, existing modeling work most often used the template-based method,[8] while no studies have ever tried ML for environmental reactions. However, ML modeling has achieved very promising results in modeling organic chemical reactions (Table 1). For example, NN models have been used to classify reaction types; Transformer and GNN models have been used to predict reaction products and retrosynthesis design.[18,32,40] Although environmental reactions and organic chemical reactions belong to two different disciplines, there are more similarities than differences from shared reaction mechanisms to commonly encountered small data set issues. The results achieved in organic reaction modeling can provide valuable insights for future environmental reaction modeling, especially regarding reaction pathways/products.

4. ML MODEL DEVELOPMENT, EVALUATION, AND INTERPRETATION

The rapid development in ML has provided diverse algorithms from basic linear correlations to more complex decision tree-based algorithms/support vector machines and to more sophisticated neural network algorithms. No single algorithm can work well for all learning tasks. Before a certain algorithm is selected, a systematic comparison among different algorithms is necessary. Readers are referred to recent comprehensive reviews for comparisons among different algorithms.[43,60] After an ML algorithm is selected and models are trained, several aspects deserve attention during the model evaluation. The first is whether the model performance is acceptable or not. For example, when using R^2 or Q^2 as the evaluation metrics for regression models, $R^2 > 0.6$ and $Q^2 > 0.5$ may be deemed acceptable.[61] The range/unit of the output may affect the interpretation of the obtained root-mean-square errors (RMSEs) when using RMSEs as the evaluation metric. For example, an RMSE of 0.28 log unit would be quite satisfactory when the adsorption coefficient (log K_d) on granular activated carbons ranged from −0.7 to 5[62] but would be less satisfactory if the output values ranged from 0 to 1. A general measure is to use multiple metrics to evaluate trained models.

The second aspect is how a new model's performance compares with that of baseline models.[33] Although chemical reaction models often achieve promising performance, they are not necessarily better than the corresponding baseline models.[63] With the prediction for the optimal reaction conditions taken as an example, one study found that the best ML model was not significantly better than naive selections that are only based on the frequency of conditions employed in the literature (e.g., selecting the most often used experimental conditions as the model prediction).[63] Also, complex models are less favorable when they only marginally outperform the corresponding simpler baseline models. As for baseline models, simple algorithms such as the K-nearest neighbor (KNN) would be a viable choice.[64,65] KNN models only use the average of K nearest neighbors in the training data set as the prediction for a test sample, so they generally do not involve complex computations and are highly efficient and easy to interpret. After the comparison with baseline models, the ML models should be further compared with traditional models or existing ML models, if available. During the comparison, it may not always be appropriate to directly compare the R^2/Q^2 or RSME values because different models are generally built on different data sets, which could considerably affect the model performance. With different data sets, other types of comparison between the new and existing models would be more appropriate, such as evaluating the model performance on external data sets that are not covered by either model[62] or evaluating the performance of the new model on the data sets used by the existing models or vice versa. Besides, it is preferable to perform new experiments or calculations to validate the newly built models. Those new results can also be added to the existing data set to further improve the model performance (details will be discussed in Section 5).[66]

In addition, it is important to point out that existing chemical reaction data sets (<2000 compounds) are generally small and biased compared to environmental chemical data sets, such as EU's Registration, Evaluation, Authorization, and Restriction of Compounds (REACH; >22 000 compounds)[67] and EPA's Distributed Structure-Searchable Toxicity (DSSTox; ≈850 000, version 2, 2019).[68] Only a small portion of the concerned compounds have been investigated due to either experimental/computational limitations or biased selections of compounds in the literature, for example, only studying the most frequently investigated or "popular" compounds.[69] These small chemical data sets may not be able to fully represent all compounds of concern; thus, the model evaluation results that rely solely on the existing data sets may not be applicable to compounds beyond the training data sets.[70] For example, a data set for modeling atmospheric reaction rate constants with ·OH mostly covers small volatile compounds.[13] The corresponding ML model or applicability domain (AD) will be inevitably limited when applied to compounds from much larger data sets such as REACH or DSSTox. Additional experiments and simulations could help address the above limitations by adding more diverse test samples.

Even when an ML model performs well on the overall test data set, the prediction may not always be satisfactory for individual test compounds, so it becomes very important to evaluate whether an ML model is applicable for certain new test compounds. In traditional QSAR modeling, the applicability domain (AD) is generally used to define the application scope of QSAR models. Commonly used AD methods include leverage, convex hull, distance-based metrics, etc.[71] However, these methods are not always suitable for ML models, for example, when the data sets have non-normal distribution[72] or the models use non-numeric representations (e.g., images or graphs) as the input. Another common approach in ML is to use chemical similarity to define the AD.[31,73] However, the obtained AD may lack model specificity; ML models for different learning tasks may have the same AD because the same chemical representation is used. Yet, different modeling tasks often rely on different types of chemical similarities. For example, the octanol−water partition coefficient (logP) is closely related to the overall chemical similarity because every part of a molecule would affect its logP, whereas the hydrogen-bonding ability depends on the similarity of certain functional groups, such as N/O-containing groups. The application of the same AD to these two different learning tasks means that the AD is not accurate/specific enough. A few studies have tried to build a separate ML model to predict the AD of another model;[74] however, the AD of the newly obtained model remains unknown. In addition, our recent study has successfully used the estimated values of widely available properties, such as logP, to conveniently calibrate estimates for properties that are less available but related, such as solute descriptors.[65]

Model interpretation is another important aspect of ML modeling. It can help understand how a prediction is obtained and whether the models follow the ground truth. Generally, there are two different approaches for model interpretation. The first is to analyze the input features ("feature-wise", right panel in Scheme 1), for example, the feature importance derived directly from algorithms like the random forest or indirectly by an external method like SHAP.[35,75] This approach can provide information about the overall importance of each input feature or the contribution of each feature toward predictions. For example, the interpretation of the model for reaction rate constants with HO· successfully identified the reactive sites and correctly classified functional groups that positively or negatively contribute to the reactivity.[35] The second approach focuses on exploring the contribution of each training sample to certain predictions ("case-wise") using the influence function method.[76] Using this approach, our recent study developed ML models for predicting solute descriptors and found that around five functionally similar training compounds were critical to obtain good predictions for a test compound.[65]

5. ML MODELING OF SMALL DATA SETS: TRANSFER LEARNING AND ACTIVE LEARNING

ML modeling has only achieved limited success in modeling environmental reaction pathways/products, possibly due to the relatively small data sets and complex reaction pathways. A small data set will make it difficult to train complex ML models and could easily lead to overfitting. Also, a small data set may lack chemical diversity and thus restrict the applicability of the built models. One way to address this is to limit the complexity of ML models, for example, reducing the depth and width of FNN models. In addition, ML studies in other fields especially in organic reactions could provide attractive solutions, such as using transfer learning (Table 2) or active learning approaches (Table 3), as detailed below.

Transfer learning is generally achieved by first training a complex ML model on a large data set and then fine-tuning it on the desired small data set.[77,78] There are two ways to employ transfer learning. First, we can use the entire or some components of an ML model from another learning task. For example, using chemical images as the input, we can employ image classification models such as ResNet[79] (Scheme 2), which is trained on natural object images,[80] to build models for reaction rate constants.[26] Second, we can add information from other models into the input of the desired model, for example, using the predictions by other models as an input for a new model.[81] The first approach has been widely adopted in various disciplines, while the second one is emerging and often requires the target task to be similar to the model to be transferred.

Despite the reported successes, we should be careful in choosing models to be transferred when applying transfer learning. For example, ResNet is trained on images of natural objects (Scheme 2A); however, chemical images usually contain short "lines" and characters to represent chemical structures. Differences between these two types of images would inevitably affect the generalization ability of the corresponding models for different learning tasks. One solution is to utilize models from closely related learning tasks, for instance, using the recent ChemNet (Scheme 2A,B)[82] rather than ResNet, because ChemNet employs chemical images or "SMILES" as the inputs and easily calculated chemical descriptors as the outputs. In a chemical transfer learning scenario, one can replace the last layer of the ChemNet with a new output layer and fine-tune new model parameters based on the target data sets. Alternatively, a recent reaction modeling study proposed another transfer learning approach to transfer more reaction-related knowledge using the entire reaction as the input. In this approach,[87] a graph neural network autoencoder (GNNAE) was first trained on a large, unlabeled reaction database. On the basis of the encoder part of the autoencoder, new models were then fine-tuned on the desired small database that contained labeled reactions. The final model considerably outperformed models that were based only on the small data set itself using traditional template-

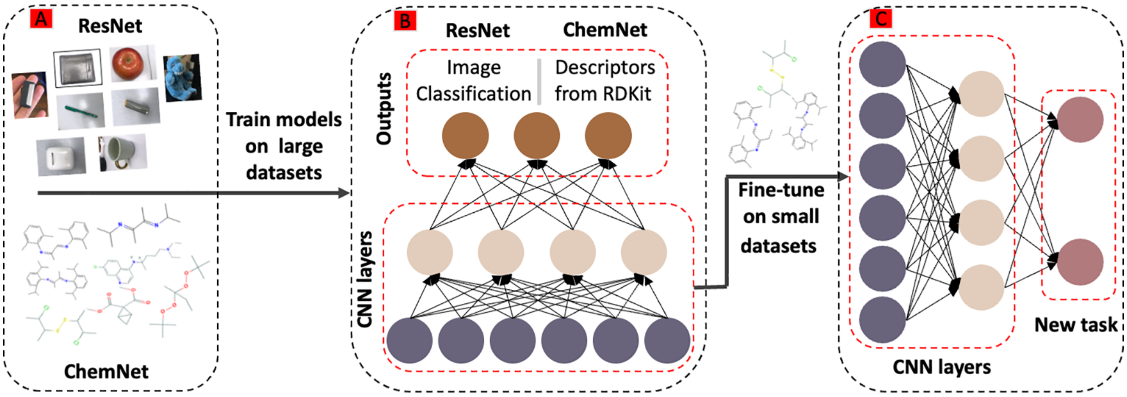
Table 2. Summary of Recent ML Modeling Using Transfer Learning

research field	learning task	input/algorithm	transfer learning	results	ref
environmental	reaction rate constants of organic pollutants with HClO, O_3, ClO_2, and $SO_4\cdot$	MF and MD/NN	use the values predicted from other models as inputs to the new ML models	RMSE = 2.04 (HClO); 1.94 (O_3); 1.49 (ClO_2); 0.7 ($SO_4\cdot$)	21
	reaction rate constants with HO·	image/CNN	fine-tune ResNet on the reaction database	RMSE decreased from (0.395−0.592) to (0.123−0.339)	26
organic chemistry or materials science	organic reaction classification	SMILES/GNN	train GNN on unlabeled reactions then fine-tune on small (4−128 per class), labeled reactions	accuracy up to 0.9 from 0.64	42
	organic reaction prediction	SMILES/Transformer	first train on a large general reaction database and then fine-tune on the desired reaction database	accuracy increased by 30% to 70%	78
	material property predictions (e.g., band gap)	stoichiometric attributes, elemental properties, and MD/LASSO	use estimated properties by traditional methods as the input to the ML models	RMSE reduced to 0.39 from 0.51 eV	81
	chemical property predictions (e.g., freesolve)	images, SMILES/CNN, and RNN	train the model first on one data set using the RDKit-calculated MDs as the outputs and then fine-tune the model for the desired chemical properties	RMSE reduced to 1.3 from 0.99	82

Table 3. ML Modeling Using Active Learning

learning task	algorithm	results	ref
virtual screening of ligands	RF + NN	accuracy of 89.3−94.8% for the top-50 000 ligands in a 100 M member library after testing only 2.4% of the candidate ligands	83
optimization of chemical reaction conditions	RF	used 0.03−0.04% of search space to finish searching; results competitive with those of human experts	84
optimization of redox potential and solubility in candidate redox couples	RF + DFT	500-fold acceleration over a random search	85
prediction of chemical/material properties	NN	10% of the data can outperform traditional models that are based on all the data	86

Scheme 2. Workflow of Transfer Learning Based on "ResNet" or "ChemNet" Using Images as the Input[a]



[a](A) Example input images of the ResNet (top) and ChemNet (bottom). (B) Model architecture for ResNet and ChemNet. The only difference between these two is the output layers: image classification for ResNet (top left) and prediction of RDKit calculated descriptors for ChemNet (top right). (C) Transfer learning: the output layer of ResNet or ChemNet is replaced with a new learning task; the CNN layers first remain unchanged and then are fine-tuned (retraining parameters of the new model on the target data set) on the input images.

based models.[87] The GNNAE approach may be favorable for environmental reaction modeling because GNNAE models can be conveniently trained using only unlabeled data (there are many more unlabeled reactions than labeled ones), which means that many more reaction data sets can be used for transfer learning. Also, as the inputs for GNNAE are the entire reactions rather than individual chemicals in the reactions, as used by other transfer learning models, more reaction information can be transferred for later modeling. As a result, major improvements might be achieved after transfer learning following the GNNAE approach.

Another useful approach is called active or adaptive learning, which essentially builds an iteration loop between ML modeling and the experiments.[83,84,88] Briefly, a base model is first developed on a small data set and then employed to make predictions for a subset of compounds from the target chemical space. Experiments are then performed to validate these predictions and to increase the training sample size to upgrade the model. Such a modeling−prediction−validation loop is often repeated several times. By such looping, active learning can quickly explore a large chemical space using a small amount of experimental data plus additional carefully designed experiments. For example, active learning has enabled researchers to identify optimal conditions for certain reactions more efficiently than human experts,[84] to achieve a 500-fold acceleration over a random search for optimal redox couples for redox flow batteries,[85] and to use a tiny amount of data but still outperform traditional methods for chemical/material property predictions (Table 3).[86] A recent study also reported a new extension of the active learning method, where rather than retraining the entire model in each iteration round, a new submodel was trained on a small number of new experimental data and then was added to the revised active learning model. When this is repeated multiple times, the final model can better utilize the small number of new data to significantly improve the model performance (e.g., about 50% more reactions were identified).[89]

The key question in active learning is how to effectively select new targets for further experiments. One study selected new targets from the chemical space where the ML models tend to make poor predictions (when compared to those of an ensemble of ML models).[86] Another study partitioned the chemical space into several subspaces and selected experimental targets from each subspace.[85] Chemical similarity can also be utilized as the criterion for the selection of experimental targets, such as always selecting compounds that are the least similar to the training data in each iteration cycle, because ML models often rely on similar training compounds to make predictions so the least similar compounds tend to have the largest prediction errors.[73] Compounds that are the least similar to the training compounds are typically located outside or on the fringe of the model's AD. Improved predictions for these compounds also suggest well-improved predictions for compounds within but near the boundary of the chemical space. When algorithms that can provide uncertainty measurements, such as the Gaussian Process and Bayesian neural networks, are used for modeling, the uncertainty associated with new predictions can also be used as a metric to select new experimental targets; that is, the predictions with higher uncertainty values will be the priority when performing the next experiments. For example, one recent study has used a Gaussian process to significantly enlarge the ADs of the models for three different chemical modeling tasks.[90] When compounds that are susceptible to poor predictions are the focus, the AD of ML models can be quickly expanded through active learning.

6. SUMMARY AND OUTLOOK

The application of ML has shed new light on chemical reaction modeling by achieving satisfactory modeling performance and expanding the input chemical representations from numerical values to strings, graphs, images, etc. However, the performance of new ML models needs to be carefully evaluated using diverse evaluation metrics and by comparing them with existing and baseline models. In addition, there are still many unresolved problems in modeling environmentally related reactions, such as reaction product modeling and data scarcity issues. Existing ML modeling for organic reactions may provide helpful insights for the first issue, and techniques such as transfer learning that have been extensively used in other fields may help address the second one. Meanwhile, it is always desirable to conduct carefully selected experiments/computations to enlarge the training data sets so that robust ML models can be built. As active learning and the robotic experiment platform technique[91] have shown promising results in effectively expanding experimental data sets in many other fields, environmental reaction modeling might greatly benefit from either or a combination of these two techniques in enlarging small data sets. It should be noted that the modeling process goes beyond the model development stage. Additional efforts should be made to check the validity of model predictions, including interpreting the model, which can help examine whether the built model violates the ground truth such as the related reaction mechanisms, obtaining the applicability domain, and using alternative measures such as the surrogate metric, which uses the accuracy in estimating widely available properties, like the octanol−water partition coefficient, to calibrate estimates for less available but related properties like Abraham descriptors.[65] If possible, the selection of a few new predictions to perform experiments would be an ideal measure to further evaluate the model.

It is also worth noting that ML and traditional models are not mutually exclusive. In fact, reaction modeling can benefit from a careful combination of both approaches. For example, one study incorporated a tree-based algorithm into the reaction profile modeling and achieved much faster modeling without sacrificing too much modeling accuracy.[92] A coupled model was successfully developed by combining FNN with the traditional adsorbed solution theory to satisfactorily predict bisolute adsorption on polymeric resins using only single-solute data.[93] Neural network models can also be used to generate the potential energy surface for molecular dynamics simulations, which can greatly facilitate the simulation without sacrificing much accuracy.[94] Such a coupling strategy could help achieve a balance between model accuracy/efficiency and interpretability.

The foundation of successful ML models is high-quality data. The current reaction databases only cover a small portion of the reported studies, so mining more reaction records from the literature should be a focus before building better models in future studies. Given more and more data being reported, it becomes critical to develop tools to facilitate or automate the data collection process. Meanwhile, many models have been developed/reported for different types of reactions. Although researchers have tried to share the built models in various ways, including source code sharing in GitHub, deployment as online prediction tools, or compilation as a standalone tool, it remains challenging for others to repeat or apply these models. Meanwhile, many studies do not share their data sets, trained models, or source codes, making it hard for others to follow these studies. A few publicly available data sets such as a recent PFAS function screen database[95] and the OQMD data set (containing quantum calculation results for over one million materials)[96] provide good examples of how to share and maintain databases. Specifically, the former database built one easily accessible framework for sharing and exploring the database of PFAS (perfluoroalkyl or polyfluoroalkyl substances), while the latter uses the widely used SQL (structural query language) database to store and manage data records. These tools are particularly helpful for handling large and ever-increasing sample sizes. Future studies may adopt these approaches or continue to develop new measures for reaction database sharing and managing, and a well-accepted protocol/method for data set/model sharing will greatly facilitate future modeling efforts.

■ 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; Email: hjz13@case.edu

Author

Kai Zhang − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States;* orcid.org/0000-0003-4058-6512

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsestwater.2c00193

Author Contributions

CRediT: **Kai Zhang** conceptualization (lead), data curation (lead), formal analysis (lead), investigation (lead), methodology (lead), validation (lead), visualization (lead), writing-original draft (lead), writing-review & editing (equal); **Huichun Zhang** conceptualization (supporting), data curation (supporting), funding acquisition (lead), investigation (supporting), methodology (supporting), project administration (lead), supervision (lead), visualization (supporting), writing-original draft (supporting), writing-review & editing (lead).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under grant # CHE-2105005.

■ GLOSSARY

Applicability Domain: The possible applicable scope of a model.

Chemical Representation: Chemical formula, structures, symbols, images, or graphs used to represent chemicals.

Data Augmentation: Increasing the amount of data by adding slightly modified copies of existing data or newly synthesized data on the basis of existing data.

Dimension Reduction: Transforming data from a high-dimensional space to a low-dimensional space so that the data can be easily visualized or used for model training.

Autoencoder: A type of model that can learn essential information from unlabeled data.

Feature: An individual property to represent part of the input.

Feature Engineering: Selecting and transforming the most relevant variables, features, from raw data when conducting modeling on the basis of domain knowledge.

Influence Function: A method to identify the contribution of a training sample to a model prediction.

Overfitting: Models follow errors, or noise, in the training data set too closely and may, therefore, fail to perform well on the test data set.

SHAP: SHapley Additive exPlanations, a method to quantify the importance of input features in individual predictions on the basis of the Shapley values.

SMILES: Simplified molecular-input line-entry system.

REFERENCES

(1) Cheng, M.; Zeng, G.; Huang, D.; Lai, C.; Xu, P.; Zhang, C.; Liu, Y. Hydroxyl radicals based advanced oxidation processes (AOPs) for remediation of soils contaminated with organic compounds: a review. *Chem. Eng. J.* **2016**, *284*, 582−598.

(2) Borch, T.; Kretzschmar, R.; Kappler, A.; Cappellen, P. V.; Ginder-Vogel, M.; Voegelin, A.; Campbell, K. Biogeochemical redox processes and their impact on contaminant dynamics. *Environ. Sci. Technol.* **2010**, *44* (1), 15−23.

(3) Zepp, R. G.; Cline, D. M. Rates of direct photolysis in aquatic environment. *Environ. Sci. Technol.* **1977**, *11* (4), 359−366.

(4) Chong, M. N.; Jin, B.; Chow, C. W.; Saint, C. Recent developments in photocatalytic water treatment technology: a review. *Water Res.* **2010**, *44* (10), 2997−3027.

(5) Mabey, W.; Mill, T. Critical review of hydrolysis of organic compounds in water under environmental conditions. *J. Phys. Chem. Ref. Data* **1978**, *7* (2), 383−415.

(6) Alexander, M. *Biodegradation and bioremediation*; Elsevier Science, 1999.

(7) Ossola, R.; Jönsson, O. M.; Moor, K.; McNeill, K. Singlet Oxygen Quantum Yields in Environmental Waters. *Chem. Rev.* **2021**, *121* (7), 4100−4146.

(8) Yuan, C.; Tebes-Stevens, C.; Weber, E. J. Reaction Library to Predict Direct Photochemical Transformation Products of Environmental Organic Contaminants in Sunlit Aquatic Systems. *Environ. Sci. Technol.* **2020**, *54* (12), 7271−7279.

(9) Zhong, S.; Zhang, H. Mn (III)-ligand complexes as a catalyst in ligand-assisted oxidation of substituted phenols by permanganate in aqueous solution. *J. Hazard. Mater.* **2020**, *384*, 121401.

(10) Huang, J.; Dai, Y.; Singewald, K.; Liu, C.-C.; Saxena, S.; Zhang, H. Effects of MnO2 of different structures on activation of peroxymonosulfate for bisphenol A degradation under acidic conditions. *Chem. Eng. J.* **2019**, *370*, 906−915.

(11) Awfa, D.; Ateia, M.; Mendoza, D.; Yoshimura, C. Application of Quantitative Structure−Property Relationship Predictive Models to Water Treatment: A Critical Review. *ACS ES&T Water* **2021**, *1* (3), 498−517.

(12) Atkinson, R. Estimation of gas-phase hydroxyl radical rate constants for organic chemicals. *Environ. Toxicol. Chem.* **1988**, *7* (6), 435−442.

(13) Kwok, E. S.; Atkinson, R. Estimation of hydroxyl radical reaction rate constants for gas-phase organic compounds using a structure-reactivity relationship: an update. *Atmos. Environ.* **1995**, *29* (14), 1685−1695.

(14) Monod, A.; Poulain, L.; Grubert, S.; Voisin, D.; Wortham, H. Kinetics of OH-initiated oxidation of oxygenated organic compounds in the aqueous phase: new rate constants, structure−activity relationships and atmospheric implications. *Atmos. Environ.* **2005**, *39* (40), 7667−7688.

(15) Gramatica, P.; Papa, E. Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure. *Environ. Sci. Technol.* **2007**, *41* (8), 2833−2839.

(16) Minakata, D.; Li, K.; Westerhoff, P.; Crittenden, J. Development of a Group Contribution Method To Predict Aqueous Phase Hydroxyl Radical (HO•) Reaction Rate Constants. *Environ. Sci. Technol.* **2009**, *43* (16), 6220−6227.

(17) Tebes-Stevens, C.; Patel, J. M.; Jones, W. J.; Weber, E. J. Prediction of Hydrolysis Products of Organic Chemicals under Environmental pH Conditions. *Environ. Sci. Technol.* **2017**, *51* (9), 5008−5016.

(18) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572−1583.

(19) Huang, K.; Zhang, H. A comprehensive kinetic model for phenol oxidation in seven advanced oxidation processes and considering the effects of halides and carbonate. *Water Research X* **2022**, *14*, 100129.

(20) Huang, J.; Jones, A.; Waite, T. D.; Chen, Y.; Huang, X.; Rosso, K. M.; Kappler, A.; Mansor, M.; Tratnyek, P. G.; Zhang, H. Fe (II) redox chemistry in the environment. *Chem. Rev.* **2021**, *121* (13), 8161−8233.

(21) 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* (1), 681−692.

(22) Xu, T.; Chen, J.; Wang, Z.; Tang, W.; Xia, D.; Fu, Z.; Xie, H. Development of Prediction Models on Base-Catalyzed Hydrolysis Kinetics of Phthalate Esters with Density Functional Theory Calculation. *Environ. Sci. Technol.* **2019**, *53* (10), 5828−5837.

(23) Meuwly, M. Machine Learning for Chemical Reactions. *Chem. Rev.* **2021**, *121* (16), 10218−10239.

(24) Lu, J.; Zhang, H.; Yu, J.; Shan, D.; Qi, J.; Chen, J.; Song, H.; Yang, M. Predicting Rate Constants of Hydroxyl Radical Reactions with Alkanes Using Machine Learning. *J. Chem. Inf. Model.* **2021**, *61* (9), 4259−4265.

(25) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55* (19), 12741−12754.

(26) Zhong, S.; Hu, J.; Yu, X.; Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chem. Eng. J.* **2021**, *408*, 127998.

(27) Cao, H.; Peng, J.; Zhou, Z.; Sun, Y.; Wang, Y.; Liang, Y. Insight into the defluorination ability of per-and polyfluoroalkyl substances based on machine learning and quantum chemical computations. *Sci. Total Environ.* **2022**, *807*, 151018.

(28) Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. py SiRC": 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* (18), 12437−12448.

(29) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466−1474.

(30) Landrum, G. *RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling*; Academic Press, 2013.

(31) 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.

(32) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10* (2), 370−377.

(33) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P.-O. Organic reactivity from mechanism to machine learning. *Nat. Rev. Chem.* **2021**, *5* (4), 240−255.

(34) Gao, Y.; Zhong, S.; Torralba-Sanchez, T. L.; Tratnyek, P. G.; Weber, E. J.; Chen, Y.; Zhang, H. Quantitative structure activity relationships (QSARs) and machine learning models for abiotic reduction of organic compounds by an aqueous Fe (II) complex. *Water Res.* **2021**, *192*, 116843.

(35) 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.* **2021**, *405*, 126627.

(36) Oldenhuis, R.; Oedzes, J. Y.; van der Waarde, J. J.; Janssen, D. B. Kinetics of chlorinated hydrocarbon degradation by Methylosinus trichosporium OB3b and toxicity of trichloroethylene. *Appl. Environ. Microbiol.* **1991**, *57* (1), 7−14.

(37) Raza, A.; Bardhan, S.; Xu, L.; Yamijala, S. S. R. K. C.; Lian, C.; Kwon, H.; Wong, B. M. A Machine Learning Approach for Predicting Defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal. *Environ. Sci. Technol. Lett.* **2019**, *6* (10), 624−629.

(38) Huang, Y.; Li, T.; Zheng, S.; Fan, L.; Su, L.; Zhao, Y.; Xie, H.-B.; Li, C. QSAR modeling for the ozonation of diverse organic compounds in water. *Sci. Total Environ.* **2020**, *715*, 136816.

(39) Borhani, T. N. G.; Saniedanesh, M.; Bagheri, M.; Lim, J. S. QSPR prediction of the hydroxyl radical rate constant of water contaminants. *Water Res.* **2016**, *98*, 344−353.

(40) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725−732.

(41) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, Long Beach, CA, USA, December 4−9, 2017; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., Garnett, R., Eds.; Curran Associates Inc.: Red Hook, NY, United States, 2017; pp 5998−6008.

(42) Wen, M.; Blau, S. M.; Xie, X.; Dwaraknath, S.; Persson, K. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chem. Sci.* **2022**, *13*, 1446−1458.

(43) Ayres, L. B.; Gomez, F. J.; Linton, J. R.; Silva, M. F.; Garcia, C. D. Taking the leap between analytical chemistry and artificial intelligence: A tutorial review. *Anal. Chim. Acta* **2021**, *1161*, 338403.

(44) Wilary, D. M.; Cole, J. M. ReactionDataExtractor: A Tool for Automated Extraction of Information from Chemical Reaction Schemes. *J. Chem. Inf. Model.* **2021**, *61* (10), 4962−4974.

(45) Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. Unassisted noise reduction of chemical reaction datasets. *Nature Machine Intelligence* **2021**, *3* (6), 485−494.

(46) Kušić, H.; Rasulev, B.; Leszczynska, D.; Leszczynski, J.; Koprivanac, N. Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: A QSAR study. *Chemosphere* **2009**, *75* (8), 1128−1134.

(47) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Commun. Math. Comput. Chem.* **2006**, *56* (2), 237−248.

(48) Gupta, S.; Basant, N. Modeling the aqueous phase reactivity of hydroxyl radical towards diverse organic micropollutants: An aid to water decontamination processes. *Chemosphere* **2017**, *185*, 1164−1172.

(49) Wang, Y.-n.; Chen, J.; Li, X.; Wang, B.; Cai, X.; Huang, L. Predicting rate constants of hydroxyl radical reactions with organic pollutants: Algorithm, validation, applicability domain, and mechanistic interpretation. *Atmos. Environ.* **2009**, *43* (5), 1131−1135.

(50) Stewart, J. J. MOPAC: a semiempirical molecular orbital program. *J. Comput. Aided Mol. Des.* **1990**, *4* (1), 1−103.

(51) Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geoscience and remote sensing letters* **2016**, *13* (3), 364−368.

(52) Cho, H.; Choi, I. S. Enhanced Deep-Learning Prediction of Molecular Properties via Augmentation of Bond Topology. *ChemMedChem.* **2019**, *14* (17), 1604−1609.

(53) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, October 22−29, 2017; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2017; pp 618−626; DOI: 10.1109/ICCV.2017.74.

(54) Jiang, D.; Wu, Z.; Hsieh, C. Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminformatics* **2021**, *13* (1), 12.

(55) Sudhakaran, S.; Amy, G. L. QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification. *Water Res.* **2013**, *47* (3), 1111−1122.

(56) Fodor, I. K. *A survey of dimension reduction techniques*; Lawrence Livermore National Laboratory: Livermore, CA, 2002.

(57) Xiao, R.; Ye, T.; Wei, Z.; Luo, S.; Yang, Z.; Spinney, R. Quantitative structure−activity relationship (QSAR) for the oxidation of trace organic contaminants by sulfate radical. *Environ. Sci. Technol.* **2015**, *49* (22), 13394−13402.

(58) Wang, Y.; Yao, H.; Zhao, S. Auto-encoder based dimensionality reduction. *Neurocomputing* **2016**, *184*, 232−242.

(59) Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.

(60) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9* (2), 513−530.

(61) Todeschini, R.; Ballabio, D.; Grisoni, F. Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *J. Chem. Inf. Model.* **2016**, *56* (10), 1905−1913.

(62) 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* (11), 7008−7018.

(63) Beker, W.; Roszak, R.; Wolos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki-Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819−4827.

(64) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure− property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 185−194.

(65) Zhang, K.; Zhang, H. Predicting Solute Descriptors for Organic Chemicals by a Deep Neural Network (DNN) Using Basic Chemical Structures and a Surrogate Metric. *Environ. Sci. Technol.* **2022**, *56* (3), 2054−2064.

(66) Gao, H.; Zhong, S.; Zhang, W.; Igou, T.; Berger, E.; Reid, E.; Zhao, Y.; Lambeth, D.; Gan, L.; Afolabi, M. A. Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization. *Environ. Sci. Technol.* **2022**, *56* (4), 2572−2581.

(67) Fantke, P.; Aurisano, N.; Provoost, J.; Karamertzanis, P. G.; Hauschild, M. Toward effective use of REACH data for science and policy. *Environ. Int.* **2020**, *135*, 105336.

(68) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology* **2019**, *12*, 100096.

(69) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **2019**, *573* (7773), 251−255.

(70) Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J. Cheminformatics* **2019**, *11* (1), 69.

(71) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52. *Alternatives to Laboratory Animals* **2005**, *33* (2), 155−173.

(72) Gajewicz, A. How to judge whether QSAR/read-across predictions can be trusted: a novel approach for establishing a model's applicability domain. *Environ. Sci. Nano* **2018**, *5* (2), 408−421.

(73) Liu, R.; Wallqvist, A. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds. *J. Chem. Inf. Model.* **2019**, *59* (1), 181−189.

(74) Sheridan, R. P. Using random forest to model the domain applicability of another random forest model. *J. Chem. Inf. Model.* **2013**, *53* (11), 2837−2850.

(75) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, Long Beach, CA, USA, December 4−9, 2017; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., Garnett, R., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp 4768−4777.

(76) Koh, P. W.; Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*; International Machine Learning Society (IMLS): Sydney, Australia, 2017; pp 1885−1894.

(77) Zhang, Y.; Wang, L.; Wang, X.; Zhang, C.; Ge, J.; Tang, J.; Su, A.; Duan, H. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. *Organic Chemistry Frontiers* **2021**, *8* (7), 1415−1423.

(78) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nat. Commun.* **2020**, *11* (1), 1−8.

(79) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, June 27−30, 2016; The Institute of Electrical and Electronics Engineers, Inc.: Piscataway, NJ, USA, 2016; pp 770−778; DOI: 10.1109/CVPR.2016.90.

(80) Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, Miami, FL, USA, June 20−25, 2009; Institute of Electrical and Electronics Engineers, Inc.: Piscataway, NJ, USA, 2009; pp 248−255; DOI: 10.1109/CVPR.2009.5206848.

(81) Zhang, Y.; Ling, C. A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials* **2018**, *4* (1), 1−8.

(82) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. Using rule-based labels for weak supervised learning: A ChemNet for transferable chemical property prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, London, United Kingdom, August 19−23, 2018; Guo, Y., Farooq, F., Eds.; Association for Computing Machinery: New York, NY, USA, 2018; pp 302−310; DOI: 10.1145/3219819.3219838.

(83) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **2021**, *12* (22), 7866−7881.

(84) Reker, D.; Hoyt, E. A.; Bernardes, G. J.; Rodrigues, T. Adaptive optimization of chemical reactions with minimal experimental information. *Cell Rep. Phys. Sci.* **2020**, *1* (11), 100247.

(85) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6* (4), 513−524.

(86) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(87) Wen, M.; Blau, S. M.; Xie, X.; Dwaraknath, S.; Persson, K. A. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chem. Sci.* **2022**, *13*, 1446−1458.

(88) Reker, D.; Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discov Today* **2015**, *20* (4), 458−465.

(89) Shim, E.; Kammeraad, J. A.; Xu, Z.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Predicting reaction conditions from limited data through active transfer learning. *Chem. Sci.* **2022**, *13* (22), 6655−6668.

(90) Zhong, S.; Lambeth, D. R.; Igou, T. K.; Chen, Y. Enlarging Applicability Domain of Quantitative Structure−Activity Relationship Models through Uncertainty-Based Active Learning. *ACS EST Engg.* **2022**, *2*, 1211.

(91) Coley, C. W.; Thomas, D. A., III; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365* (6453), 1−9.

(92) Keller, C. A.; Evans, M. J. Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10. *Geoscientific Model Development* **2019**, *12* (3), 1209−1225.

(93) Zhang, K.; Zhang, H. Coupling a Feedforward Network (FN) Model to Real Adsorbed Solution Theory (RAST) to Improve Prediction of Bisolute Adsorption on Resins. *Environ. Sci. Technol.* **2020**, *54* (23), 15385−15394.

(94) Zeng, J.; Cao, L.; Xu, M.; Zhu, T.; Zhang, J. Z. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **2020**, *11* (1), 1−9.

(95) Su, A.; Rajan, K. A database framework for rapid screening of structure-function relationships in PFAS chemistry. *Scientific Data* **2021**, *8* (1), 1−10.

(96) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **2015**, *1* (1), 1−15.