Environmental Science & Technology

pubs.acs.org/est

Feature

Machine Learning: New Ideas and Tools in Environmental Science and Engineering

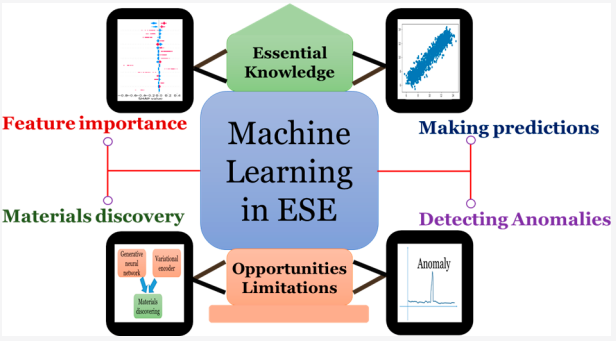
Shifa Zhong, Kai Zhang, Majid Bagheri, Joel G. Burken, April Gu, Baikun Li, Xingmao Ma, Babetta L. Marrone, Zhiyong Jason Ren, Joshua Schrier, Wei Shi, Haoyue Tan, Tianbao Wang, Xu Wang, Bryan M. Wong, Xusheng Xiao, Xiong Yu, Jun-Jie Zhu, and Huichun Zhang*

Cite This: *Environ. Sci. Technol.* 2021, 55, 12741−12754

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The rapid increase in both the quantity and complexity of data that are being generated daily in the field of environmental science and engineering (ESE) demands accompanied advancement in data analytics. Advanced data analysis approaches, such as machine learning (ML), have become indispensable tools for revealing hidden patterns or deducing correlations for which conventional analytical methods face limitations or challenges. However, ML concepts and practices have not been widely utilized by researchers in ESE. This feature explores the potential of ML to revolutionize data analysis and modeling in the ESE field, and covers the essential knowledge needed for such applications. First, we use five examples to illustrate how ML addresses complex ESE problems. We then summarize four major types of applications of ML in ESE: making predictions; extracting feature importance; detecting anomalies; and discovering new materials or chemicals. Next, we introduce the essential knowledge required and current shortcomings in ML applications in ESE, with a focus on three important but often overlooked components when applying ML: correct model development, proper model interpretation, and sound applicability analysis. Finally, we discuss challenges and future opportunities in the application of ML tools in ESE to highlight the potential of ML in this field.



KEYWORDS: *applicability domain, artificial intelligence, best practices, feature importance, machine learning modeling, model applications, model interpretation, predictive modeling*

1. INTRODUCTION

The rapid advancement in environmental analytical tools and monitoring technologies has led to a corresponding explosive expansion in both the quantity and complexity in data generation, which demands more advanced and powerful computational and data analytical approaches beyond traditional statistical tools. Data analytical approaches that have less dependence on prior knowledge, such as machine learning (ML), have shown promise in solving complex data patterns or formats because of their powerful fitting abilities. As a result, the past decade has witnessed a rapid growth of ML, especially deep learning, in a variety of applications, such as image classification and machine translation. These tools are revolutionizing many scientific fields, from chemistry,[1] material sciences,[2] and biomedicine,[3] to quantum physics.[4] Researchers in the broad field of environmental science and engineering (ESE) have also adopted ML enthusiastically, as demonstrated by the explosive growth in the number of publications (5855 between 1990 and 2020) on the applications of ML in ESE (Figure 1). These applications cover broad areas, including assessing environmental risks, evaluating the health of water and wastewater infrastructure, optimizing treatment technologies, identifying and characterizing pollution sources, and performing life cycle analysis, among others.

The definition of ML is that "ML algorithms build a model based on sample data, known as 'training data', to make predictions or decisions without being explicitly programmed to do so".[5] Example ML algorithms include random forest, support vector machines, and artificial neural networks.[6−8] Deep learning is one class of ML, in which "deep" refers to the multilayered neural network structures,[9] such as recurrent neural networks and convolution neural networks.[10−12] Every ML algorithm can be decomposed into three key components: the structure of the algorithm, such as random forest and deep

Published: August 17, 2021

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

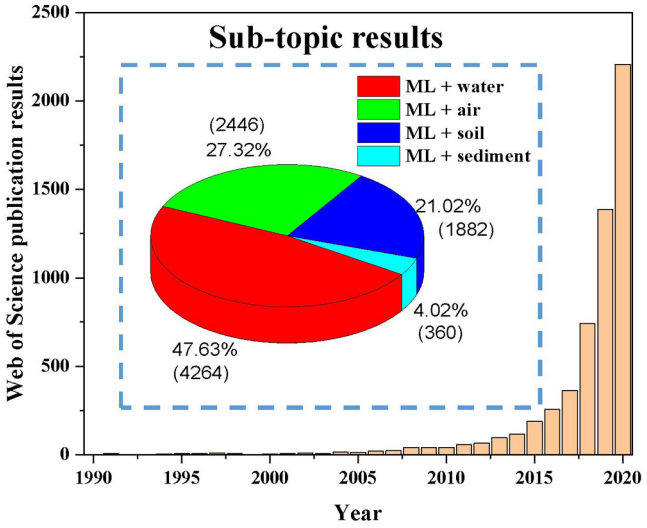


Figure 1. Number of publications in ML applications in ESE based on the Web of Science (access date 1/28/2021) with the keyword "machine learning" combined with the categories of environmental science, water resources, public environmental occupational health, environmental engineering, and environmental studies. The inset shows the subtopic results from 1990 to 2020 with the keywords specified.

neural networks; the goal to achieve, such as prediction accuracy and squared error; and the training method to achieve the goal, such as stochastic gradient descent.[13] One major advantage of ML algorithms is that they can easily identify trends or patterns in data without human intervention. Their predictive performance can be continually improved with more available data. Multidimensional and multivariety data can also be handled by ML even in dynamic or uncertain environments.[14−17]

Many complex ESE problems can be addressed by ML. Five specific examples are explained in section 2, and a comprehensive summary is included in section 3. A challenge, and opportunity, is that even as ML algorithms are increasingly applied to ESE problems, these algorithms are being continuously improved and new ones are developed. ESE practitioners passively benefit from increasingly advanced ML methods, as they allow more challenging problems to be addressed. ESE practitioners can also make an active contribution to the development of more advanced ML methods by identifying and communicating the encountered limitations to ML researchers.

Despite the initial success of ML in ESE, many concerns remain. First, ESE researchers may be eager to use ML in their research but lack the knowledge of how to properly employ it, which may lead to incorrect applications of ML to their data sets. Second, increased data volume and complexity has enabled the use of more advanced ML algorithms, such as deep neural networks, capable of capturing sophisticated nonlinear relationships. However, these types of models are often of a "black box" nature; therefore, model interpretation is vital for investigating whether ML model predictions are consistent with the fundamental principles of the domain science. Although there is a growing field of model interpretability,[18] such interpretation is still commonly neglected in ESE.[19−22] Alternatively, emerging approaches to building neural networks that enforce physical symmetries or directly implement differential equations that parallel chemical kinetic differential equations,[23,24] have not yet been adopted in ESE. Third, applicability domain (AD) analysis of ML models is yet to be practiced by researchers in ESE after model development,[19−22] other than development of quantitative structure−activity relationships (QSARs).[25] To introduce ML to the general ESE researcher's toolbox, this feature aims to discuss the current status, essential knowledge, shortcomings, challenges, and future opportunities of ML in ESE to highlight the potential of ML in the ESE field.

2. HOW DOES ML ADDRESS ESE PROBLEMS?

ESE applications often utilize supervised ML approaches, which are essentially interpolation—given a big enough library of inputs (independent variables or "features") matched with outputs (dependent variables or "outcomes"), these algorithms after training can take a new input and predict its corresponding output. Where this differs from traditional statistical tools is the ability to treat a large number of features that have weak or nonlinear relationships with the outcomes. ML can be more effective than traditional statistical tools in handling various data formats, such as text, images, and graphs, where the important information is not contained in a single input variable, nor are the important variables known ahead of time, but instead where some previously unknown combination of features are needed to determine the outcome. These unique properties of ML are especially suitable for solving complex environmental problems with rich sets of input features.[26] We, here, take five specific examples from different environmental fields to illustrate how ML can address complex environmental problems. We will have a comprehensive discussion of current ML applications in ESE in section 3.

2.1. Prediction of Particulate Matter ($PM_{2.5}$). When predicting daily to yearly variations in $PM_{2.5}$ for a region, the major drivers include meteorological conditions, such as air temperature, dew point temperature, visibility, pressure, potential evaporation, downward longwave radiation flux, downward shortwave radiation flux, relative humidity, and wind; and land-use variables, such as limited access highway, highway, local roads, and forest cover.[27] In addition, there are regional differences associated with meteorological conditions and daily variations in $PM_{2.5}$ levels. These factors interact with each other to form complex relationships, which are challenging to process using traditional statistical tools but can be effectively handled by ML when sufficient training data are given to learn their relationships. For example, whereas multiple linear regression of PM2.5 prediction from these factors only achieves an R^2 of 0.60, an artificial neural network using the same features and data can achieve a higher R^2 of 0.74.[28]

2.2. Prediction of Water Resource Availability. Climate change—increasing average temperature, uneven shifting of precipitation patterns, and frequent occurrence of extreme climate events like floods and droughts—can substantially impact the predictability and availability of water resources across regions. To better manage and sustain future water resources, it is essential to develop decision support tools that can handle these variations and uncertainties, which arise due to either interactions between natural and human systems or from the variability of climate. Addressing such challenges requires insights into past and future patterns, yet this insight can be difficult to develop based on traditional statistical approaches. In contrast, ML allows effective analysis and

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

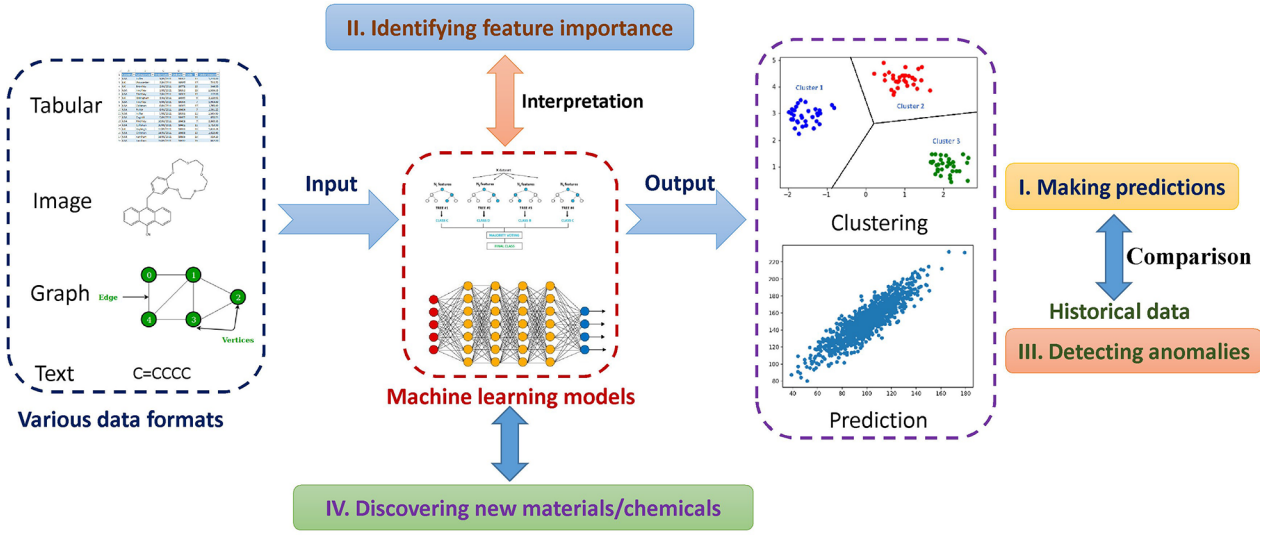


Figure 2. Four common applications of ML in ESE. Different data formats can be used for inputs in ML to develop models for (I) prediction purposes. By interpreting the ML models, the importance of features (II) can be obtained. By comparing the prediction with the historical data, (III) anomaly detection can be achieved. ML can also be used to (IV) discover new materials.

prediction of future water availability through processing data related to climate change and water system interactions.[29]

2.3. Data Collection and Interpretation Across Water Facilities. An enormous amount and wide spectrum of data in water utilities have been collected from in-place supervisory control and data acquisition systems, including flow rate, temperature, dissolved oxygen concentration, turbidity, and chlorine content, to name a few, as well as data from ex situ laboratory information management systems and computerized maintenance management systems. Although these rich and complex data sets have been around for years, the gathered data sets from treatment facilities are often incomplete and not connected to each other. The current approach to data collection, interpretation, and utilization is not suitable for rapid identification of malfunction, swift control and adjustment under transient fluctuations, or efficient decision-making regarding facility operations. This is because these traditional models are mainly based on statistics,[30−34] which are valid only for a limited operating range and cannot capture the time-varying or nonlinear behavior of dynamic systems. In contrast, ML models can adapt to fast-changing situations and, because they do not rely on predetermined rules, they can use varied, dynamic data to update themselves for better predictions.[35,36]

2.4. Modeling of Biochemical Wastewater Treatment Systems. Most wastewater treatment plants (WWTPs) use activated sludge and other processes to remove contaminants, primarily including organic carbon, nitrogen, and phosphorus. There can be tens of thousands of different microbial species present in each system. The abundance of different organisms, as well as the time-varying and highly nonlinear characteristics of the WWTP process, makes the relationship with process performance a difficult one to investigate. Deterministic models based on the fundamental biokinetics for activated sludge processes are not particularly practical due to the complex nature of biological reactions, highly multivariable aspects of treatment plants, the expertise required to calibrate activated sludge models, and system-specific calibration.[37] The influent wastewater flow and composition also vary over time and follow dynamic patterns. However, ML techniques could predict sludge bulking in wastewater treatment plants with improved accuracies without the burden of calibration.[38] ML carries the potential to characterize the relationships between microbial communities and parameters of the process based on the fusion of sequencing data and model outputs. For example, Zhu et al. compared the abilities of nine single or hybrid models to predict the next day's influent wastewater flow rate.[39] Compared to principal component regression (adjusted $R^2 \approx 0.66$; mean absolute percentage error $\approx$ 18.2%) and partial least-squares (0.70; 17.5%), the hybrid model using an artificial neural network (0.83; 13.3%) achieved the best results.

2.5. Prediction and Identification of Endocrine-Disrupting Chemicals (EDCs). Currently, there are over one hundred thousand chemicals available on the market, many of which lack toxicity data, including data on their endocrine disrupting activity.[40] In screening and predicting potential EDCs, many traditional tools, such as read-cross, are based on the concept that structurally similar compounds carry similar toxicity and these tools have been successfully applied.[41,42] However, focusing on a limited set of chemicals or only structurally similar compounds makes it difficult to apply these traditional tools to tens of thousands of untested chemicals. Additionally, because of the complex molecular mechanisms behind endocrine disruption, many traditional tools that rely solely on structural information usually experience "activity cliffs", that is, a group of seemingly structurally similar chemicals have different endocrine disrupting activities.[43,44] These problems stress the unsuitability of traditional data analysis methods, and researchers have tried many approaches to combine ML tools with big data (including various biological, physicochemical information, etc.) to solve them. For instance, a computational method was developed to automatically extract useful biologically active characteristic fragments of the estrogen receptor alpha (ERα) and the androgen receptor (AR) based on more than 4000 chemicals and to built chemical fragment−in vitro relationships (118 for ERα and 99 for AR) to illustrate new molecular mechanisms of EDCs.[45] Pathway activity prediction, such as the AR-mediated pathway, used 1746 compounds from 11 high-throughput in vitro screening assays, considered multiple

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

points (receptor binding, coregulator recruitment, gene transcription, and protein production) and multiple cell types, and showed extremely high prediction accuracy.[46,47] Furthermore, combined with 16 different ML methods, pathway activity information was used to build 91 predictive models for EDCs' activity on the AR. The final consensus model showed a highly predictive performance on 87 914 chemicals.[47] Similar research also can be found in an ER study that used 19 different ML methods and 32 464 chemical data sets to build 48 models for ER activity prediction.[48]

3. CURRENT STATUS OF ML APPLICATIONS IN ESE

Figure 2 shows four common applications of ML in ESE: making predictions, identifying feature importance, detecting anomalies, and discovering new materials or chemicals. Table 1 provides more detailed examples for each of the applications.

3.1. Making Predictions. Overall, making predictions is one of the most popular applications of ML in ESE. This application can be implemented through either regression or classification modeling. The key assumption is that the distribution of training examples provided to the algorithm is representative of the examples that the model will be asked to predict. In this sense, it does not matter if the problem is time-dependent or -independent, as long as "past" observations are predictive of "future" queries. In addition, if the predictions are related to the location or sites, they can be seen as geographic space-based predictions, such as predicting arsenic concentrations in groundwater at different sites.[49] Most prediction applications are in the realm of supervised learning, in which the sample outputs are labeled. For example, ML methods have been extensively employed to predict changes of various wastewater variables, such as nitrogen, phosphorus, solids, chemical oxygen demand (COD), biochemical oxygen demand (BOD), and future flow rate;[39,50−53] or atmospheric pollutants, such as $PM_{2.5}$ and carbonaceous aerosols.[54−56]

In addition to supervised learning, unsupervised ML approaches have also emerged recently. In unsupervised methods, the computer algorithm works on its own to find patterns in the data, that is, in an "unsupervised" fashion without intervention from the user. Most applications of unsupervised ML have been used to automatically categorize data into separate groups or "clusters" that have similar characteristics. In the context of environmental studies, there have been recent unsupervised ML applications, such as *t*-distributed stochastic neighbor embedding (t-SNE) or *k*-means clustering, on categorizing the carbon−fluorine bond dissociation energies of per- and polyfluoroalkyl substances (PFAS) to understand bond dissociation energies.[57,58] These algorithms allow the visualization of high-dimensional data as two-dimensional "clusters" where data points grouped within a cluster share similar characteristics with each other. It is important to reiterate that these clusters were automatically chosen by these unsupervised ML algorithms, without human intervention. As such, these results demonstrate that these algorithms can be useful tools for automatically classifying and rationalizing chemical trends in environmental contaminants that would otherwise have been difficult to detect manually.

3.2. Identifying Feature Importance. This refers to the techniques of assigning scores to input features or independent predictors to evaluate their relative importance to the outcome. This is also referred to as "model interpretability," "feature extraction", or the study of "latent spaces" in the ML literature.[18] For example, air pollution, as a complex global environmental issue, is affected by many factors. ML techniques such as support vector machines, neural networks, and feature extraction methods can be especially useful to determine the most significant factors for the modeling of particulate matter.[27] The prerequisite for model interpretation is that the obtained ML model has satisfactory predictive performance or ML has already correctly "learned" the underlying relationship between the features and the outcome. Thus, interpretation reveals the implicit knowledge "learned" by the ML model and indicates whether the model is based on a correct "understanding" of the underlying mechanisms, which is important for validating the model. This process may also yield new knowledge. For example, Mori et al. interpreted ML models to provide evidence-based information on how stressors and ecologically important environmental factors interact and drive ecological processes and microbial biomass.[98] Such information cannot be obtained without performing model interpretation. Similarly, there are several attempts to interpret ML models to correlate factors with the chemical activity of EDCs, including unveiling features that make EDCs chemically active, determining the type of activity on the ERα or the AR, and how these features exert their functions.[45] Using ML techniques, researchers also found that the octanol−water partitioning coefficient (log K_{ow}) plays a dominant role in regulating plant uptake of organic contaminants, while their molecular weight plays a secondary role.[99] In recent years, data-driven analytics such as ML have become key tools for discovery in public health and ESE research to find hidden patterns and causal relationships and to identify key features, that is, chemical exposure or other social economic parameters, that are linked with health outcomes.[84,100]

3.3. Anomaly Detection. This refers to the identification of historical or current abnormal events to avoid irregularities or unreliable operations; this is widely used to prevent credit card fraud.[101] The basic principle is that new observations are compared to the learned distribution of (mostly)-normal historical data to determine statistically improbable deviations. For instance, anomaly detection has been used to identify burst locations of pipes and detect contamination events in water distribution networks,[90−92] with the former contributing to less water loss and the latter important for reducing public health risks. ML can also be used to make predictions for a future event by comparing those predictions with current data to identify potential outliers and then calculate the probability of future contamination events.

3.4. Discovering Materials and Chemicals. Discovering materials and chemicals based on ML is another rapidly growing application in ESE, such as designing environmentally friendly adsorbents and catalysts. Take biopolymers as an example, researchers are developing biodegradable polymer materials that can functionally replace plastics made from fossil feedstocks, thus reducing plastic pollution in the environment. Toward this goal, they apply an innovative ML approach to bring together two adaptive codesign loops: (i) a chemical loop, where the structure and function of possible chemical combinations of the polymer backbone and side chains are explored for their predicted properties[96,102] and (ii) a synthetic biology adaptive design loop, in which the biosynthetic pathways involved in polymer production and the roles of their component genes and proteins, are investigated and targeted for improvement by genetic engineering methods. Another example is to develop novel adsorbents, where two

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

Table 1. Example Applications of ML in ESE

application	media	example	input	output	performance	refs
making predictions	air	$PM_{2.5}$ formation	meteorological conditions, land-use variables, etc.	$PM_{2.5}$ concentration	better than multiple linear models	27, 28, 59
		prediction of carbonaceous aerosols	air quality variables and meteorological conditions	organic carbon, elemental carbon concentrations	better than generalized additive models	56
	water	QSARs for predicting the reactivity of contaminants	molecular fingerprints or molecular descriptors	reaction rate constants	better than multiple linear regression	60−63
		single- and multisolute adsorption modeling	descriptors for chemicals and adsorbents	adsorbed amounts or coefficients	outperformed multilinear regression models	64, 65
		deep learning and systematic understanding of urban water management infrastructure	water quantity and quality	energy consumption, chemical use, and environmental impacts	understood systems complexity, identified unnecessary energy/chemical inputs, and reduced unwanted effects	66
		predicting water quality parameters for wastewater effluent	various operating parameters such as mixed liquor suspended solids	water quality characteristics	outperformed activated sludge models	67−69
		intelligent monitoring of membrane fouling in water and wastewater treatment	operating parameters and influent water characteristics	membrane permeability	introduced an integrated approach to control membrane fouling	70, 71
		monitoring of water or wastewater quality and risk management	water/wastewater quality variables	monitoring variables	superior over multiple linear regression model	51
		prediction and early warning of combined sewer and facility overflow	rainfall data in conjunction with meteorological conditions	future combined sewage flow rate	better than classical, statistical methods	39
		predicting arsenic concentrations in groundwater	various climate and soil parameters, geology, and topography	probability of arsenic in groundwater exceeding the World Health Organization (WHO) guideline	better than statistical models	49
	soil/sediment	prediction of plant uptake efficiency	physicochemical properties of chemicals	concentration factors	improved accuracy compared to mechanistic models	72
		neural network predictive models for cadmium and cerium uptake by plants	plant properties	cadmium and cerium concentrations	showed high accuracy and low error	73
	toxicity	QSARs for predicting toxic end points of high concern, e.g., bioconcentration factor, skin sensitization, carcinogenicity, mutagenicity, developmental toxicity	"SMILES" of chemicals, molecular descriptors, and physicochemical properties, relevant experimental results	qualitative and quantitative activity prediction results	improved the prediction accuracy and innovatively predicted some in vivo toxicity end points with complex mechanisms that traditional tools could not predict	74−77
		identifying potential endocrine-disrupting chemicals (EDCs)	1D, 2D, and 3D molecular descriptors, physicochemical properties, in vitro and in vivo assays	EDC activity	active pollutants can be prioritized for regulators	47, 48, 78−81
identifying feature importance	air	understanding the relationship between pollutant concentrations and influencing factors	meteorological and land-use variables	$PM_{2.5}$	identified the most important variables to $PM_{2.5}$ prediction	27
	water	understanding the relationship between factors and the EDCs' chemical activity (chemical features)	physicochemical properties of chemicals, high throughput screening of in vitro data (e.g., ToxCast from U.S.EPA), and toxicity data in vivo	active/inactive and the types of toxic activity	showed high prediction accuracy, broke the limitation of the "black box", and identified structural features	45, 47, 82
	soil/sediment	understanding the relationship between factors and uptake/transport of organic contaminants	physicochemical properties of chemicals	uptake efficiency	introduced new relationships for plant uptake of contaminants	83
	environmental health	identifying key causal factors that predict environmental and human health outcomes	environmental exposure, water quality parameters, social economic factors	public health outcome (i.e., premature birth rate)	improved the identification of multiple factors that contribute to and predict health outcome	84, 85
	others	understanding the molecular mechanisms between environmental pollutants and nuclear receptors	computational information, such as molecular docking and classical molecular dynamics simulations	ligand−receptor interactions	found key residues and binding modes, built new virtual screening models	86, 87
		mechanism profiling of cellular stress response, hepatotoxicity caused by oxidative stress	in vitro assay models, time-series toxicogenomic data and big data	ranking in vitro−in vivo correlations; most relevant bioassay(s) related to hepatotoxicity, identifying key molecular toxicity mechanisms	fully explored the source-to outcome continuum of modern experimental toxicology	88, 89
		reducing animal experiments and exploring the relationship between in vitro bioassays and in vivo acute toxicity end points	high throughput screening in vitro data and in vivo acute toxicity data	nonanimal models for extrapolating in vitro to in vivo	built chemical fragment−in vitro−in vivo relationships	43

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

Table 1. continued

application	media	example	input	output	performance	refs
detecting anomalies	water	identifying contamination events	water quality indicators, for example, chlorine, electric conductivity, pH, temperature, total organic carbon, turbidity	probability of contamination	improved the identification of the source of fecal pollution in water	90, 91
		identifying the burst location of pipes	flow, pressure, or demand	probability of burst	deep learning models can successfully predict the bursts in a real-life network	92
		identifying abnormal events in wastewater treatment processes	wastewater influent and process variables	classified wastewater variables	successfully detected instrument and process anomalies	93
	chemical assessment	QSAR models for governments, industry, and other stakeholders to fill gaps in (eco)toxicity data needed for assessing chemical hazards	chemical structural information and environmental information	structural characteristics and potential mechanism or mode of action of a target chemical	many QSAR models have been widely accepted, such as the OECD (Organization for Economic Co-operation and Development) QSAR Toolbox	94
discovering new materials or chemicals	air	discovering new adsorption materials for CO_2	edges, vertices, topologies of metal–organic frameworks (MOFs)	new MOF structures	discovered MOFs are strongly competitive against some of the best-performing MOFs/zeolites ever reported	95
	water	discovering and designing new bioplastics materials, with desired durability and faster degradability in the environment than conventional plastics	polymer property data from literature and experiments, polymer features based on composition, configuration, topology, and polarity	prediction of target properties (e.g., Glass Transition Temperature) from candidate polymers	relationships determined between polymer chemical compositions and target properties; enabled the prediction of thermomechanical performance	96
	others	drug discovery and development of environmentally friendly chemicals; enabling new environmental technology	systematic and comprehensive high-dimensional data, single cell technology data	new environmentally friendly compounds; determine microbial community diversity and sampling size	challenges lie primarily with the lack of interpretability and repeatability of ML-generated results	97

different strategies have been used: the generative adversarial network (e.g., to discover new zeolite structures)[103] and the variational autoencoder (e.g., to obtain new structures for a metal−organic framework).[104] Such research typically has three steps: (1) train a model based on one of the ML algorithms, (2) generate new structures using the generator or decoder, and (3) validate the new structures by experiments or molecular simulation.

4. ESSENTIAL KNOWLEDGE REQUIREMENT AND CURRENT SHORTCOMINGS IN ML APPLICATIONS IN ESE

Although ML has shown great potential for solving ESE issues, inexperience with ML may lead to unsatisfactory performance or inappropriate applications of ML tools. This problem has been recognized by the closely affiliated material science and chemistry communities, which have felt the need to address "best practices" for ML.[105,106] For example, applying cross-validation to the entire data set is one of the most common mistakes made by inexperienced researchers. Cross-validation should be applied to the training data set to screen ML algorithms and tune the hyperparameters,[107] and a completely independent test data set should be reserved to test the predictive performance of the obtained model.[108,109] Researchers may confuse the validation data set with the test data set.[110] Not doing this properly may have serious overfitting risks.[111] Seeking collaborations between researchers in the two fields can resolve these issues, but ESE researchers should acquire basic knowledge to fully benefit from the advantages of ML. For example, feature engineering (e.g., which features to use) and hyperparameters (e.g., the depth of trees in a random-forest model or the number of layers in a deep neural network) play important roles in the design of ML models, but often draw upon domain expertise. Therefore, effective collaboration will only be possible if ESE researchers understand fundamental ML concepts.

Figure 3 shows a typical workflow of ML model development and highlights key steps that should be carefully followed to deliver meaningful results when applying ML to ESE. The benefits of ML tools can only be realized if they are correctly used. For example, a common assumption in ML applications is that a larger volume of data will always generate better predictive performance of the ML model. However, more "representative" data rather than "big" data are more important for obtaining robust, powerful ML models. The term "representative" refers to the diversity of data. If a data set has too many similar data points, that is, they are close to each other in mathematical space, even though the sample size is large the model will not necessarily have the desired predictive performance. For instance, collecting chemicals with diverse chemical structures is preferred for developing widely applicable QSAR models. Once ML is deemed appropriate for a given application, collecting a sufficient amount of "representative" data is vital to the subsequent model development. It is also important that all key input features are considered in the input, as stated in a recent article that *"if a well-trained neural network provided with sufficient data cannot achieve the desired accuracy, it is likely that the data provided is not entirely relevant to the task being predicted."*[112] The same statement may be true for other ML algorithms. Therefore, a deep understanding of the ESE application is crucial for developing correct and meaningful ML models.

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

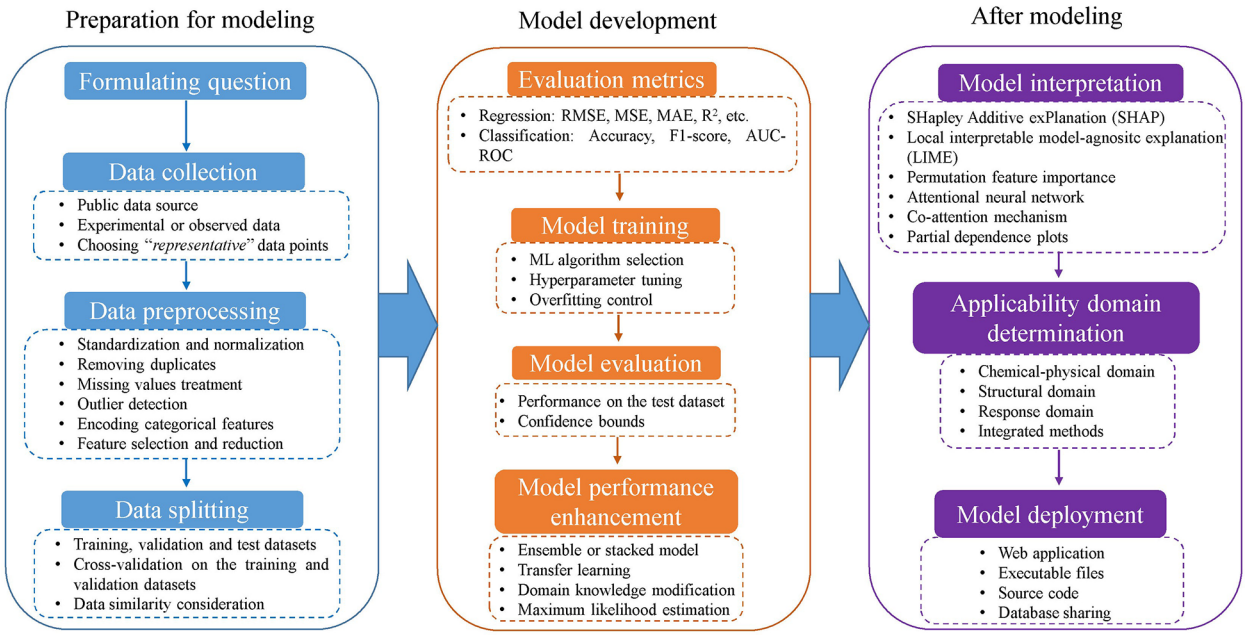


Figure 3. Typical workflow for the development of ML models, where RMSE, MSE, MAE, and R^2 refer to the root mean squared error, mean squared error, mean absolute error, and coefficient determination, while AUC-ROC means the area under the curve−receiver operating characteristic curve. For more details about specific ML techniques listed in Figure 3, please refer to the following references: data preprocessing, data splitting, and model development;[113,114] model interpretation, including SHAP,[115] LIME,[116] permutation feature importance,[6] attentional neural network,[117] coattention mechanism,[118] and partial dependence plots;[119] and applicability domain determination.[25]

Once a data set is ready, proper data preprocessing and splitting are necessary to avoid possible data leakage. Data leakage refers to cases where information used in model testing is inadvertently present during the model training process, so that the estimated prediction performance on the test set overestimates the true predictive performance of the model.[120] For example, duplicate samples may result in the same sample in both the training and test data sets, and the model may simply memorize these data without generalizing to new data points. Data leakage problems can often be subtle. To avoid data leakage, data standardization or normalization should be conducted based on the training data set and then applied to the test data set. For data splitting, the test data set should always be set aside and never be used for model development to ensure the independence of the test data set in evaluating the model performance. Cross-validation should be applied to the training data set rather than the entire data set, and the purpose is not to deliver a final predictive model but to evaluate the unbiased generalization ability of the ML algorithms. Feature selection can improve prediction accuracy, reduce model training time, and reduce the overfitting risk. Better model performance is commonly observed when an effective feature selection method is applied.[39,50]

During model training, different evaluation metrics are used in different applications to evaluate the model performance; but in general, multiple metrics should be employed for a given model. For an imbalanced data set in which the number of positive cases is more than the number of negative cases, accuracy cannot be a reliable evaluation metric. Instead, metrics such as the F1 score,[121] AUC (area under the curve)−ROC (receiver operating characteristics),[122] or Matthews correlation coefficient (MCC) should be employed.[123] Tuning hyperparameters is required to calibrate a successful model.[107,124] The optimum hyperparameters are determined to be the ones that minimize the errors on the validation data set. After model development, the predictive performance of the model is evaluated on the test data set, which should never be used in model training or hyperparameter selection. This is consistent with a real application where the query sample is "new" and not merely identical with a previously observed training example. The model performance might be further improved by developing ensemble or stacked models, applying transfer learning, or performing domain knowledge modification. For example, Xiao et al. has developed an ensemble model that combines random forest, a generalized additive model, and extreme gradient boosting to predict historical $PM_{2.5}$ concentrations in China based on satellite data.[22] However, it is not guaranteed that ensemble models are always better than single models, for example, when predicting the concentrations of ultrafine particles in the air.[19]

Model interpretation is an indispensable step after model development because it examines if the model predictions are consistent with the domain science (i.e., ground truth). Certain ML algorithms (e.g., ensemble methods, neural networks) are less interpretable than traditional statistical models.[125,126] It will be problematic if a model with apparently high predictive performance makes predictions that are physically impossible; therefore, it is sometimes necessary to adjust the model and "teach" the model the correct domain science to obtain results that make sense. For example, based on the interpretation of the results of an ML model, one may need to remove the incorrectly identified features or adjust their weights and bias and retrain the model. Additional benefits related to model interpretation include (1) validating the reliability of the model, that is, if the model makes predictions aligned with domain knowledge; (2) unveiling implicit knowledge or even new knowledge that the model reveals; (3) informing feature engineering, such as selecting the most useful features; and (4) informing data collection, such as selecting data containing the desirable features during data collection.

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

Thus far, model interpretation has been largely overlooked in ESE. An example from the biomedical field demonstrates the necessity of model interpretation. McCloskey et al. used the attribution method to interpret their model and discovered that a deep neural network still learns spurious binding logic, despite its perfect classification accuracy on the protein−ligand binding data set.[127] Zhong et al. recently used the Shapley method to interpret how a deep neural network model and an extreme gradient boosting model made predictions on rate constants toward hydroxyl radicals (log $k_{OH^{\bullet}}$).[60] They found that the former model assigned equal contributions of the same atom groups to log $k_{OH^{\bullet}}$ no matter which compounds this atom group was in, which contradicts the chemical principle that the reactivity of a structural functional group is affected by its chemical environment. By comparison, the latter model correctly considered these contributions differently. Moreover, there are many complex ESE issues for which identifying the dominant drivers is extremely difficult. In these cases, if a successful model is obtained, one may be able to extract some implicit new knowledge that is otherwise unrecognized (see section 3.2 for more details).

Applicability domain (AD) is a concept from QSARs. It defines a domain in which the activity of a new compound can be reliably predicted based on the calibrated QSAR.[25] Recently this concept has been applied to materials science problems.[128] After model development, verifying the reliability of a prediction is the key idea of applicability domain analysis. However, studies seldom conduct AD analyses except for those applying QSAR models.[25] We propose expanding this concept to other ML models after they have been successfully developed and validated. There is no unified method to determine the AD of a model because diverse data types are used, such as image, text, and tabular data; and several measures may be applicable, for example, chemical-physical or structural domains for QSAR models.[25] Nevertheless, AD should be defined with a clear mathematic form, such as the similarity calculation for QSAR models.[25]

Model deployment means how your model is shared with others. Commonly used model deployment approaches include sharing source code, providing executable files, and web applications. Sharing source code allows others to reuse or modify the models but requires expertise in coding; web applications or executable files provide ready-to-use tools to make predictions but limit the ability of other researchers to modify or augment the tools. Deploying ML models in more than one way may help ML models reach more users.

5. CHALLENGES

As an emerging tool for ESE, ML faces many challenges in applying it successfully. The following challenges are listed based on their order in the model development process, from data collection to model application.

5.1. Data Scarcity and Quality. The first challenge is how to effectively collect valid, high-quality data. For instance, since the outbreak of the COVID-19 pandemic, over 20 000 peer-reviewed papers have been published in a few months, among which many efforts have been launched to detect this virus in raw and treated wastewater.[129−134] Ironically, the almost unlimited accessibility of data makes it hard to obtain the most relevant and qualified data. No generally agreed upon guidelines are available to sample, transport, store, and analyze the wastewater samples, leading to variations in the data quality. These low-quality data cannot be easily used for modeling. On the other hand, ML requires a large sample size to build robust predictive models and make accurate predictions.[135−137] Unfortunately, there is often a lack of mature databases for numerous environmental applications, and the data are often either scattered or not available in the literature. For example, researchers have systematically summarized the modes of action (MOAs) for thyroid hormone receptors (TRs)-related endocrine disruptions.[138] Developing high-throughput screening (HTS) in vitro assays based on MOAs can quickly generate a large volume of valid and reliable data for ML modeling. However, until recently, few in vitro assays were available to examine MOAs. For instance, EDCs can interact with one of 26 molecular processes to induce TRs-related endocrine disruptions, but only 14 out of the 26 (53.8%) processes have assays that have demonstrated reliability and are available for screening endocrine disrupting potentials.[138] Even when there are data, different researchers often conduct experiments under different conditions of water quality, soil/sediment properties, catalyst or adsorbent type and loading, etc., which creates discrepancies among the collected data. To develop predictive models that are truly robust and widely applicable, we should first build a large, consistent source data set.

5.2. Overfitting. Overfitting means that a model shows excellent predictive performance on the training samples but fails to accurately predict for new samples. This is especially a problem for more complex ML algorithms which contain large numbers of learned parameters. Concerns about overfitting and a lack of interpretability hinder, for example, the further optimization of automatic control processes in water/wastewater systems via ML algorithms compared with first-principle theoretical models.[139] Detecting overfitting remains a challenge, but methods that can reduce the overfitting risk include feature selection,[140] data augmentation,[141] cross-validation,[108] regularization,[142] model simplification/choice,[143] dropout,[144] and early stopping.[145]

5.3. Bias of ML Models. In a technical sense, bias refers to systematic distortions of data sets and algorithms that result in undesirable outcomes. Data set biases occur when the training data are not representative of the planned use case and can arise when training data are inadvertently contaminated with the desired outcome information or when the training data are missing relevant examples. Data contamination problems are a more subtle form of the data leakage discussed above, and can be difficult to identify without domain expertise. For example, ML models trained to predict the severity of pneumonia from patient X-rays can inadvertently learn to distinguish the type of machine used to acquire the image (for example, the sickest patients are scanned with a portable X-ray machine, whereas healthier patients can be moved to a more sophisticated machine), rather than the information in the intended image itself.[146] Missing data is a problem in scientific communities, as peer-reviewed publications only report the most promising positive results, omitting negative results and outliers, which are crucial for ML model performance.[147] Data can also be missing because of choices by human scientists, such as habits in using particular reagents, reaction conditions, or sampling plans, or neglecting to collect data that violate a currently accepted theory. These types of anthropogenic biases in data sets also degrade ML performance.[148] Algorithmic bias occurs when the inherent structure of the model or loss function does not correspond to the desired use case. For example, the application of a linear classifier to data with nonlinear concave

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

or convex boundaries cannot succeed; it is necessary to choose a different type of ML model capable of describing the complex relationship embedded in the data. Similarly, the use of an inappropriate success metric (discussed above) during the training process will result in unintended predictions. The timely identification of the possible bias in an ML model is crucial for its applications (beyond that of even environmental issues). The issue of bias can be mitigated by improving the interpretability of the ML model, where domain knowledge can be incorporated to institute judgment on its validity. A practical strategy to identify bias in ML models is via convening an ensemble of ML models, which compares the results of different ML models on the same set of problems. This comparison will help identify the consistency in ML model performance and allow bias associated with a particular ML model to be identified.

5.4. Other Underlying Concerns. Besides the above-mentioned challenges, there are still some broader concerns that deserve to be further discussed. (1) We should not over trust or overestimate ML tools. It is always necessary to verify the findings either experimentally or based on the domain knowledge or experience. (2) Traditional statistical tools may be more appropriate than ML in some cases, such as when there are small sample sizes. (3) Not every ESE problem can be solved by ML tools directly. How to elegantly convert such ESE problems to ones that can be addressed by ML requires artful design.

6. FUTURE OPPORTUNITIES AND OUTLOOK

Although challenges lie ahead, there are still many opportunities, as described below.

6.1. Balance Model Fidelity and Interpretability. Environmental systems are characterized by complex interactions of various parameters and processes. Conventional modeling is often based on significant simplifications and assumptions relying on the domain knowledge of human experts. Despite many challenges to adopting ML in ESE, ML offers a unique opportunity to precisely predict the input and output relationships for complicated systems without a priori hypotheses. ML models are designed to make the most accurate predictions possible by understanding the complex relationships embedded in data. However, the complexity of environmental problems leads to an added layer of challenge regarding interpretation and bias for ML. The interpretation of the modeling results is often difficult as ML models are like "black box" networks in which the impact of inputs on the output cannot be understood easily. Integration of an ML model with a traditional mechanistic model may retain their respective merits and can be used to solve comprehensive problems. We encourage ESE researchers to think creatively within their domains and use ML when appropriate.

6.2. Data Sharing. Data sharing should be a consensus in ESE because a significant amount of time is spent on collecting and cleaning data before model development. It is essential to build an open-access data-sharing community in ESE where we can help increase the size and diversity of data by adding more experimental or observed data points, similar to ImageNet (http://www.image-net.org/). In this way, more accurate models with larger applicability domains can be developed to better serve ESE. It is desirable for the obtained databases and source codes for the models to be released to the public through platforms, such as GitHub (http://github.com/), a widely used web-based service for code sharing, review, and management of open-source projects. Repositories such as DLHub (http://dlhub.org) not only archive data and program code, but also make it easy to use the models without the need to install software. We recommend that journal editors and peer-reviewers require that papers relying upon ML provide complete data sets and analysis source code in an electronic format (either in public repositories or as supporting information) as a necessary condition for publication, unless there are compelling privacy or legal nondisclosure reasons that preclude data-sharing.

6.3. Data Collection from Trusted Sources. Similar to other research areas, it is challenging to be inclusive of all valid publications in ESE and extract necessary information from different sources. Data can be obtained from stakeholders (e.g., water and wastewater utilities, government databases, etc.) or directly measured using instruments. Another trusted source of big data is the vast amount of literature available from scientific articles and reports. These traditional data collections are usually based on structured and homogeneous data, while textual data from the literature consider unstructured and structured texts with different formats and types using linguistic and statistical techniques. Literature text mining is a new area for ML and a void in ESE. Using ML methods to analyze ESE literature data effectively can reveal trends and patterns and underlying knowledge in complex and ever-changing environmental systems. For example, Zhu et al. proposed novel methods that analyze the complex publication data across many domains using deep text preprocessing,[149] such as acronym and abbreviation detection, chemical expression identification, and synonym combination, which has been beneficial in extracting the true value of big literature data. We recommend the creation and adoption of formal data and metadata schemas by the ESE community to facilitate data reuse.

6.4. Applications of ML Models. Beyond providing accurate predictions, one next step would be to create additional knowledge; for example, new materials or applications. On the one hand, ML models can help "discover" new information from the existing data. For example, one research group identified eight antibacterial compounds that are structurally distant from known antibiotics from the ZINC15 database (>107 million chemicals) based on an ML model.[150] On the other hand, ML can help create new chemical spaces beyond manually defined heuristics, which enables exploration of a vast landscape of possible material chemistries and properties that would not be possible to explore with an Edisonian trial and error approach.[151] The flexibility of adjustment and the capability of pseudoreal time calculation and validation under varying conditions uniquely enable ML to have precise control of diverse ESE systems. This use of ML thus facilitates discovery and innovation.

6.5. Educational Opportunities. To take advantage of the rapid development of ML and unparalleled computational power, it is imperative that the next generation of ESE practitioners are prepared to properly utilize these tools. We recommend the creation of interdisciplinary certificates, specializations, and degrees in, for example, *Environmental Data Science*, which will train the next generation of environmental engineers and scientists at the interface of computer, data, and environmental sciences to prepare them for these challenges. Environmental students in this track may take several essential computer and data science courses, such as *Data Structures, Data Science, Intro to AI, Machine Learning,*

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

and *Programming*. Computer and data science students may take courses from the ESE curriculum to receive a certificate or minor in ESE.

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

Authors

Shifa Zhong − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States*

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

Majid Bagheri − *Department of Civil, Architectural, and Environmental Engineering, Missouri University of Science and Technology, Rolla, Missouri 65409, United States*

Joel G. Burken − *Department of Civil, Architectural, and Environmental Engineering, Missouri University of Science and Technology, Rolla, Missouri 65409, United States;* orcid.org/0000-0002-7774-5364

April Gu − *Department of Civil and Environmental Engineering, Cornell University, Ithaca, New York 14850, United States;* orcid.org/0000-0002-5099-5531

Baikun Li − *Department of Civil and Environmental Engineering, University of Connecticut, Storrs, Connecticut 06269, United States;* orcid.org/0000-0002-5623-5912

Xingmao Ma − *Department of Civil and Environmental Engineering, Texas A&M University, College Station, Texas 77843, United States*

Babetta L. Marrone − *Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States*

Zhiyong Jason Ren − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0001-7606-0331

Joshua Schrier − *Department of Chemistry, Fordham University, The Bronx, New York 10458, United States;* orcid.org/0000-0002-2071-1657

Wei Shi − *School of Environment, Nanjing University, Nanjing 210093, China*

Haoyue Tan − *School of Environment, Nanjing University, Nanjing 210093, China*

Tianbao Wang − *Department of Civil and Environmental Engineering, University of Connecticut, Storrs, Connecticut 06269, United States*

Xu Wang − *School of Civil and Environmental Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China; Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China;* orcid.org/0000-0003-4555-1108

Bryan M. Wong − *Department of Chemical & Environmental Engineering, Materials Science & Engineering Program, University of California-Riverside, Riverside, California 92521, United States;* orcid.org/0000-0002-3477-8043

Xusheng Xiao − *Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, Ohio 44106, United States*

Xiong Yu − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States*

Jun-Jie Zhu − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0002-7546-2870

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c01339

Notes

The authors declare no competing financial interest.

Biography

Huichun Zhang is the Frank H. Neff Professor in the Department of Civil and Environmental Engineering at Case Western Reserve University (US). She received her B.S. (1994) and M.S. (1997) in environmental chemistry from Nanjing University (China) and her Ph.D. (2004) in Environmental Engineering from Georgia Institute of Technology (US). Her major research interests include interfacial reduction−oxidation processes in complex mixtures and the fate and transformation of emerging organic contaminants in natural and engineered environments. Her recent research areas also include predictive modeling for contaminant reactivity and sorption using both classical models and machine learning tools.

ACKNOWLEDGMENTS

H.Z. acknowledges the financial support by the National Science Foundation under Grant CBET-1804708. The authors are thankful for helpful suggestions from Dr. Ming Hu at the Cleveland Clinic Foundation, Cleveland, OH, and Dr. Zhen Cheng at Shanghai Jiao Tong University, China. B.L.M. acknowledges financial support from the Laboratory Directed Research and Development (LDRD) program of the Los Alamos National Laboratory (LANL) via project (#20190001DR).

REFERENCES

(1) Janet, J. P.; Kulk, H. J. *Machine Learning in Chemistry*; American Chemical Society, 2020; p 1.

(2) Selvaratnam, B.; Koodali, R. T. Machine learning in experimental materials chemistry. *Catal. Today* **2021**, *371*, 77−84.

(3) Goecks, J.; Jalili, V.; Heiser, L. M.; Gray, J. W. How Machine Learning Will Transform Biomedicine. *Cell* **2020**, *181* (1), 92−101.

(4) Dunjko, V.; Briegel, H. J. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Rep. Prog. Phys.* **2018**, *81* (7), 074001.

(5) Koza, J. R.; Bennett, F. H.; Andre, D.; Keane, M. A., Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In *Artificial Intelligence in Design '96*,

Gero, J. S., Sudweeks, F.,, Eds.; Springer Netherlands: Dordrecht, 1996; pp 151−170.
(6) Breiman, L. Random forests. *Machine learning* **2001**, *45* (1), 5−32.
(7) Joachims, T. Svmlight: Support vector machine. *SVM-Light Support Vector Machine*, 1999. http://svmlight.joachims.org/.
(8) Wang, S.-C. Artificial neural network. In *Interdisciplinary Computing in Java Programming*; Springer, 2003; pp 81−100.
(9) Deng, L.; Yu, D. Deep Learning: Methods and Applications. *FNT in Signal Processing* **2013**, *7* (3−4), 197−387.
(10) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, *61*, 85−117.
(11) Traore, B. B.; Kamsu-Foguem, B.; Tangara, F. Deep convolution neural network for image recognition. *Ecological Informatics* **2018**, *48*, 257−268.
(12) Sak, H.; Senior, A.; Beaufays, F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv*, 2014, 1402.1128. https://arxiv.org/abs/1402.1128.
(13) Bottou, L. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*; Springer, 2012; pp 421−436.
(14) Ikonomakis, M.; Kotsiantis, S.; Tampakas, V. Text classification using machine learning techniques. *WSEAS Transactions on Computers* **2005**, *4* (8), 966−974.
(15) Loussaief, S.; Abdelkrim, A. In *Machine Learning Framework for Image Classification*, 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2016; IEEE, 2016; pp 58−61.
(16) Ahn, W.-Y.; Ramesh, D.; Moeller, F. G.; Vassileva, J. Utility of machine-learning approaches to identify behavioral markers for substance use disorders: impulsivity dimensions as predictors of current cocaine dependence. *Frontiers in Psychiatry* **2016**, DOI: 10.3389/fpsyt.2016.00034.
(17) Brunton, S. L.; Kutz, J. N. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*; Cambridge University Press, 2019.
(18) Molnar, C. *Interpretable Machine Learning*. Lulu.com, 2020.
(19) Kerckhoffs, J.; Hoek, G.; Portengen, L. t.; Brunekreef, B.; Vermeulen, R. C. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. *Environ. Sci. Technol.* **2019**, *53* (3), 1413−1421.
(20) Pak, U.; Ma, J.; Ryu, U.; Ryom, K.; Juhyok, U.; Pak, K.; Pak, C. Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Sci. Total Environ.* **2020**, *699*, 133561.
(21) Song, R.; Keller, A. A.; Suh, S. Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. *Environ. Sci. Technol.* **2017**, *51* (18), 10777−10785.
(22) Xiao, Q.; Chang, H. H.; Geng, G.; Liu, Y. An ensemble machine-learning model to predict historical PM2. 5 concentrations in China from satellite data. *Environ. Sci. Technol.* **2018**, *52* (22), 13260−13269.
(23) Mattheakis, M.; Protopapas, P.; Sondak, D.; Di Giovanni, M.; Kaxiras, E. Physical symmetries embedded in neural networks. *arXiv*, 2019, 1904.08991. https://arxiv.org/abs/1904.08991.
(24) Ji, W.; Deng, S. Autonomous discovery of unknown reaction pathways from data by chemical reaction neural network. *J. Phys. Chem. A* **2021**, *125* (4), 1082−1092.
(25) Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability domain for QSAR models: where theory meets reality. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)* **2016**, *1* (1), 45−63.
(26) Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **2018**, *15* (4), 233−234.
(27) Hu, X.; Belle, J. H.; Meng, X.; Wildani, A.; Waller, L. A.; Strickland, M. J.; Liu, Y. Estimating PM2. 5 concentrations in the conterminous United States using the random forest approach. *Environ. Sci. Technol.* **2017**, *51* (12), 6936−6944.
(28) Gupta, P.; Christopher, S. A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. *J. Geophys. Res.* **2009**, *114* (D20), D20205.
(29) Wongso, E.; Nateghi, R.; Zaitchik, B.; Quiring, S.; Kumar, R. A Data-Driven Framework to Characterize State-Level Water Use in the United States. *Water Resour. Res.* **2020**, *56* (9), e2019WR024894.
(30) Harrou, F.; Dairi, A.; Sun, Y.; Senouci, M. Statistical monitoring of a wastewater treatment plant: A case study. *J. Environ. Manage.* **2018**, *223*, 807−814.
(31) Wang, X.; Ratnaweera, H.; Holm, J. A.; Olsbu, V. Statistical monitoring and dynamic simulation of a wastewater treatment plant: A combined approach to achieve model predictive control. *J. Environ. Manage.* **2017**, *193*, 1−7.
(32) Yu, Y.; Zou, Z.; Wang, S. Statistical regression modeling for energy consumption in wastewater treatment. *J. Environ. Sci.* **2019**, *75*, 201−208.
(33) Gernaey, K. V.; van Loosdrecht, M. C.; Henze, M.; Lind, M.; Jørgensen, S. B. Activated sludge wastewater treatment plant modelling and simulation: state of the art. *Environ. Model. Software* **2004**, *19* (9), 763−783.
(34) GarcÇa-Alvarez, D. In *Fault Detection Using Principal Component Analysis (PCA) in a Wastewater Treatment Plant (WWTP)*, Proceedings of the International Student's Scientific Conference, 2009; 2009; pp 55−60.
(35) Guo, H.; Jeong, K.; Lim, J.; Jo, J.; Kim, Y. M.; Park, J.-p.; Kim, J. H.; Cho, K. H. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *J. Environ. Sci.* **2015**, *32*, 90−101.
(36) Granata, F.; de Marinis, G. Machine learning methods for wastewater hydraulics. *Flow Meas. Instrum.* **2017**, *57*, 1−9.
(37) Mannina, G.; Cosenza, A.; Viviani, G. Uncertainty assessment of a model for biological nitrogen and phosphorus removal: Application to a large wastewater treatment plant. *Physics and Chemistry of the Earth, Parts A/B/C* **2012**, *42*, 61−69.
(38) Bagheri, M.; Mirbagheri, S. A.; Bagheri, Z.; Kamarkhani, A. M. Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach. *Process Saf. Environ. Prot.* **2015**, *95*, 12−25.
(39) Zhu, J.-J.; Anderson, P. R. Performance evaluation of the ISMLR package for predicting the next day's influent wastewater flowrate at Kirie WRP. *Water Sci. Technol.* **2019**, *80* (4), 695−706.
(40) Hartung, T.; Rovida, C. Chemical regulators have overreached. *Nature* **2009**, *460* (7259), 1080−1081.
(41) Patlewicz, G.; Ball, N.; Becker, R. A.; Booth, E. D.; Cronin, M. T.; Kroese, D.; Steup, D.; Van Ravenzwaay, B.; Hartung, T. Food for thought: read-across approaches−misconceptions, promises and challenges ahead. *Alternatives to Animal Experimentation: ALTEX* **2014**, *31* (4), 387−396.
(42) Wang, N. C. Y.; Zhao, Q. J.; Wesselkamper, S. C.; Lambert, J. C.; Petersen, D.; Hess-Wilson, J. K. Application of computational toxicological approaches in human health risk assessment. I. A tiered surrogate approach. *Regul. Toxicol. Pharmacol.* **2012**, *63* (1), 10−19.
(43) Russo, D. P.; Strickland, J.; Karmaus, A. L.; Wang, W.; Shende, S.; Hartung, T.; Aleksunes, L. M.; Zhu, H. Nonanimal models for acute toxicity evaluations: applying data-driven profiling and read-across. *Environ. Health Perspect.* **2019**, *127* (4), 047001.
(44) Maggiora, G. M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535−1535.
(45) Tan, H.; Wang, X.; Hong, H.; Benfenati, E.; Giesy, J. P.; Gini, G. C.; Kusko, R.; Zhang, X.; Yu, H.; Shi, W. Structures of Endocrine-Disrupting Chemicals Determine Binding to and Activation of the Estrogen Receptor α and Androgen Receptor. *Environ. Sci. Technol.* **2020**, *54* (18), 11424−11433.
(46) Kleinstreuer, N. C.; Ceger, P.; Watt, E. D.; Martin, M.; Houck, K.; Browne, P.; Thomas, R. S.; Casey, W. M.; Dix, D. J.; Allen, D.; Sakamuru, S.; Xia, M.; Huang, R.; Judson, R. Development and

Validation of a Computational Model for Androgen Receptor Activity. *Chem. Res. Toxicol.* **2017**, *30* (4), 946−964.
(47) Mansouri, K.; Kleinstreuer, N.; Abdelaziz, A. M.; Alberga, D.; Alves, V. M.; Andersson, P. L.; Andrade, C. H.; Bai, F.; Balabin, I.; Ballabio, D.; et al. CoMPARA: collaborative modeling project for androgen receptor activity. *Environ. Health Perspect.* **2020**, *128* (2), 027002.
(48) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenati, E.; Muratov, E.; Wedebye, E.; Grisoni, F.; Mangiatordi, G. F.; Incisivo, G. M.; Hong, H.; Ng, H. W.; Tetko, I. V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; Nikolov, N. G.; Nicolotti, O.; Andersson, P. L.; Zang, Q.; Politi, R.; Beger, R. D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S. A.; Slavov, S.; Hu, X.; Judson, R. S. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016**, *124* (7), 1023−1033.
(49) Podgorski, J.; Berg, M. Global threat of arsenic in groundwater. *Science* **2020**, *368* (6493), 845−850.
(50) Haimi, H.; Mulas, M.; Corona, F.; Vahala, R. Data-derived soft-sensors for biological wastewater treatment plants: An overview. *Environ. Model. Software* **2013**, *47*, 88−107.
(51) Zhu, J.-J.; Kang, L.; Anderson, P. R. Predicting influent biochemical oxygen demand: Balancing energy demand and risk management. *Water Res.* **2018**, *128*, 304−313.
(52) Newhart, K. B.; Holloway, R. W.; Hering, A. S.; Cath, T. Y. Data-driven performance analyses of wastewater treatment plants: A review. *Water Res.* **2019**, *157*, 498−513.
(53) Wang, Z.; Man, Y.; Hu, Y.; Li, J.; Hong, M.; Cui, P. A deep learning based dynamic COD prediction model for urban sewage. *Environmental Science: Water Research & Technology* **2019**, *5* (12), 2210−2218.
(54) Wang, J.; Song, G. A deep spatial-temporal ensemble model for air quality prediction. *Neurocomputing* **2018**, *314*, 198−206.
(55) Ma, J.; Cheng, J. C.; Lin, C.; Tan, Y.; Zhang, J. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmos. Environ.* **2019**, *214*, 116885.
(56) Zhu, J.-J.; Chen, Y.-C.; Shie, R.-H.; Liu, Z.-S.; Hsu, C.-Y. Predicting carbonaceous aerosols and identifying their source contribution with advanced approaches. *Chemosphere* **2021**, *266*, 128966.
(57) 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.
(58) Su, A.; Rajan, K. A database framework for rapid screening of structure-function relationships in PFAS chemistry. *Sci. Data* **2021**, *8* (1), 14.
(59) Lyu, B.; Hu, Y.; Zhang, W.; Du, Y.; Luo, B.; Sun, X.; Sun, Z.; Deng, Z.; Wang, X.; Liu, J.; et al. Fusion Method Combining Ground-Level Observations with Chemical Transport Model Predictions Using an Ensemble Deep Learning Framework: Application in China to Estimate Spatiotemporally-Resolved PM2. 5 Exposure Fields in 2014−2017. *Environ. Sci. Technol.* **2019**, *53* (13), 7306−7315.
(60) 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.
(61) 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.
(62) 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.
(63) 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.
(64) 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.
(65) 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.
(66) Wang, X.; Wang, Z.; Pan, Y.; Luo, Y.; Liu, J.; Yang, M. Perspective and Prospects on Applying Artificial Intelligence to Address Water and Environmental Challenges of 21st Century. *Bulletin of Chinese Academy of Sciences* **2020**, *35* (9), 1163−1176.
(67) Bagheri, M.; Mirbagheri, S. A.; Ehteshami, M.; Bagheri, Z.; Kamarkhani, A. M. Analysis of variables affecting mixed liquor volatile suspended solids and prediction of effluent quality parameters in a real wastewater treatment plant. *Desalin. Water Treat.* **2016**, *57* (45), 21377−21390.
(68) Mirbagheri, S. A.; Bagheri, M.; Boudaghpour, S.; Ehteshami, M.; Bagheri, Z. Performance evaluation and modeling of a submerged membrane bioreactor treating combined municipal and industrial wastewater using radial basis function artificial neural networks. *Journal of Environmental Health Science and Engineering* **2015**, *13* (1), 17.
(69) Mokhtari, H. A.; Bagheri, M.; Mirbagheri, S. A.; Akbari, A. Performance evaluation and modelling of an integrated municipal wastewater treatment system using neural networks. *Water Environ. J.* **2020**, *34*, 622−634.
(70) Bagheri, M.; Akbari, A.; Mirbagheri, S. A. Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning techniques: A critical review. *Process Saf. Environ. Prot.* **2019**, *123*, 229−252.
(71) Mirbagheri, S. A.; Bagheri, M.; Bagheri, Z.; Kamarkhani, A. M. Evaluation and prediction of membrane fouling in a submerged membrane bioreactor with simultaneous upward and downward aeration using artificial neural network-genetic algorithm. *Process Saf. Environ. Prot.* **2015**, *96*, 111−124.
(72) Bagheri, M.; Al-jabery, K.; Wunsch, D.; Burken, J. G. Examining plant uptake and translocation of emerging contaminants using machine learning: Implications to food security. *Sci. Total Environ.* **2020**, *698*, 133999.
(73) Rossi, L.; Bagheri, M.; Zhang, W.; Chen, Z.; Burken, J. G.; Ma, X. Using artificial neural network to investigate physiological changes and cerium oxide nanoparticles and cadmium uptake by *Brassica napus* plants. *Environ. Pollut.* **2019**, *246*, 381−389.
(74) Cassano, A.; Manganaro, A.; Martin, T.; Young, D.; Piclin, N.; Pintore, M.; Bigoni, D.; Benfenati, E. In CAESAR models for developmental toxicity, *Chemistry Central Journal*, Springer: 2010; pp 1-11
(75) Zhu, H.; Rusyn, I.; Richard, A.; Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure−activity relationship models of animal carcinogenicity. *Environ. Health Perspect.* **2008**, *116* (4), 506−513.
(76) Zhu, H.; Ye, L.; Richard, A.; Golbraikh, A.; Wright, F. A.; Rusyn, I.; Tropsha, A. A novel two-step hierarchical quantitative structure−activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environ. Health Perspect.* **2009**, *117* (8), 1257−1264.
(77) Rahman, S. M.; Lan, J.; Kaeli, D.; Dy, J.; Alshawabkeh, A.; Gu, A. Z., Machine Learning-based Biomarkers Identification and Validation from Toxicogenomics-Bridging to Regulatory Relevant Phenotypic Endpoints. *bioRxiv*, 2020, DOI: 10.1101/2020.12.18.423486.

(78) Wang, X.; Zhang, X.; Xia, P.; Zhang, J.; Wang, Y.; Zhang, R.; Giesy, J. P.; Shi, W.; Yu, H. A high-throughput, computational system to predict if environmental contaminants can bind to human nuclear receptors. *Sci. Total Environ.* **2017**, *576*, 609−616.
(79) Rotroff, D. M.; Dix, D. J.; Houck, K. A.; Knudsen, T. B.; Martin, M. T.; McLaurin, K. W.; Reif, D. M.; Crofton, K. M.; Singh, A. V.; Xia, M.; et al. Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals. *Environ. Health Perspect.* **2013**, *121* (1), 7−14.
(80) Rotroff, D. M.; Martin, M. T.; Dix, D. J.; Filer, D. L.; Houck, K. A.; Knudsen, T. B.; Sipes, N. S.; Reif, D. M.; Xia, M.; Huang, R.; et al. Predictive endocrine testing in the 21st century using in vitro assays of estrogen receptor signaling responses. *Environ. Sci. Technol.* **2014**, *48* (15), 8706−8716.
(81) Reif, D. M.; Martin, M. T.; Tan, S. W.; Houck, K. A.; Judson, R. S.; Richard, A. M.; Knudsen, T. B.; Dix, D. J.; Kavlock, R. J. Endocrine profiling and prioritization of environmental chemicals using ToxCast data. *Environ. Health Perspect.* **2010**, *118* (12), 1714−1720.
(82) Chen, Q.; Wang, X.; Shi, W.; Yu, H.; Zhang, X.; Giesy, J. P. Identification of thyroid hormone disruptors among Ho-PBDEs: in vitro investigations and coregulator involved simulations. *Environ. Sci. Technol.* **2016**, *50* (22), 12429−12438.
(83) Bagheri, M.; Al-jabery, K.; Wunsch, D. C.; Burken, J. G. A deeper look at plant uptake of environmental contaminants using intelligent approaches. *Sci. Total Environ.* **2019**, *651*, 561−569.
(84) Dong, S.; Feric, Z.; Li, X.; Rahman, S. M.; Li, G.; Wu, C.; Gu, A. Z.; Dy, J.; Kaeli, D.; Meeker, J. In *A Hybrid Approach to Identifying Key Factors in Environmental Health Studies*, 2018 IEEE International Conference on Big Data (Big Data), 2018; IEEE, 2018; pp 2855−2862.
(85) Dong, S.; Feric, Z.; Li, G.; Wu, C.; Gu, A. Z.; Dy, J.; Meeker, J.; Padilla, I. Y.; Cordero, J.; Vega, C. V. Using Undersampling with Ensemble Learning to Identify Factors Contributing to Preterm Birth. *arXiv*, 2020, 2009.11242. https://arxiv.org/abs/2009.11242.
(86) Kolšek, K.; Mavri, J.; Sollner Dolenc, M.; Gobec, S.; Turk, S. Endocrine Disruptome—An Open Source Prediction Tool for Assessing Endocrine Disruption Potential through Nuclear Receptor Binding. *J. Chem. Inf. Model.* **2014**, *54* (4), 1254−1267.
(87) Kortagere, S.; Krasowski, M. D.; Reschly, E. J.; Venkatesh, M.; Mani, S.; Ekins, S. Evaluation of computational docking to identify pregnane X receptor agonists in the ToxCast database. *Environ. Health Perspect.* **2010**, *118* (10), 1412−1417.
(88) Kim, M. T.; Huang, R.; Sedykh, A.; Wang, W.; Xia, M.; Zhu, H. Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant response element reporter gene assay models and big data. *Environ. Health Perspect.* **2016**, *124* (5), 634−641.
(89) Gao, C.; Weisman, D.; Gou, N.; Ilyin, V.; Gu, A. Z. Analyzing high dimensional toxicogenomic data using consensus clustering. *Environ. Sci. Technol.* **2012**, *46* (15), 8413−8421.
(90) Housh, M.; Ostfeld, A. An integrated logit model for contamination event detection in water distribution systems. *Water Res.* **2015**, *75*, 210−223.
(91) Ballesté, E.; Belanche-Muñoz, L. A.; Farnleitner, A. H.; Linke, R.; Sommer, R.; Santos, R.; Monteiro, S.; Maunula, L.; Oristo, S.; Tiehm, A.; et al. Improving the identification of the source of faecal pollution in water using a modelling approach: From multi-source to aged and diluted samples. *Water Res.* **2020**, *171*, 115392.
(92) Zhou, X.; Tang, Z.; Xu, W.; Meng, F.; Chu, X.; Xin, K.; Fu, G. Deep learning identifies accurate burst locations in water distribution networks. *Water Res.* **2019**, *166*, 115058.
(93) Dairi, A.; Cheng, T.; Harrou, F.; Sun, Y.; Leiknes, T. Deep learning approach for sustainable WWTP operation: A case study on data-driven influent conditions monitoring. *Sustainable Cities and Society* **2019**, *50*, 101670.
(94) Bhatia, S.; Schultz, T.; Roberts, D.; Shen, J.; Kromidas, L.; Marie Api, A. Comparison of Cramer classification between Toxtree, the OECD QSAR Toolbox and expert judgment. *Regul. Toxicol. Pharmacol.* **2015**, *71* (1), 52−62.
(95) Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence* **2021**, *3* (1), 76−86.
(96) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model.* **2019**, *59* (12), 5013−5025.
(97) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* **2019**, *18* (6), 463−477.
(98) Mori, N.; Debeljak, B.; Škerjanec, M.; Simčič, T.; Kanduč, T.; Brancelj, A. Modelling the effects of multiple stressors on respiration and microbial biomass in the hyporheic zone using decision trees. *Water Res.* **2019**, *149*, 9−20.
(99) Bagheri, M.; Al-Jabery, K.; Wunsch, D.; Burken, J. G. Examining plant uptake and translocation of emerging contaminants using machine learning: Implications to food security. *Sci. Total Environ.* **2020**, *698*, 133999.
(100) Dong, Y.; Xu, L.; Yang, Z.; Zheng, H.; Chen, L. Aggravation of reactive nitrogen flow driven by human production and consumption in Guangzhou City China. *Nat. Commun.* **2020**, *11* (1), 1209.
(101) Kou, Y.; Lu, C.-T.; Sirwongwattana, S.; Huang, Y.-P. In *Survey of Fraud Detection Techniques*, IEEE International Conference on Networking, Sensing and Control, 2004; IEEE, 2004; pp 749−754.
(102) Bejagam, K. K.; Iverson, C. N.; Marrone, B. L.; Pilania, G. Molecular dynamics simulations for glass transition temperature predictions of polyhydroxyalkanoate biopolymers. *Phys. Chem. Chem. Phys.* **2020**, *22* (32), 17880−17889.
(103) Kim, B.; Lee, S.; Kim, J. Inverse design of porous materials using artificial neural networks. *Science Advances* **2020**, *6* (1), eaax9324.
(104) Yao, Z.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O.; Snurr, R. Q. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nature Machine Intelligence* **2021**, *3*, 76−86.
(105) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine learning for materials scientists: An introductory guide toward best practices. *Chem. Mater.* **2020**, *32* (12), 4954−4965.
(106) Vishwakarma, G.; Sonpal, A.; Hachmann, J. Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry. *Trends in Chemistry* **2021**, *3* (2), 146−156.
(107) Claesen, M.; De Moor, B., Hyperparameter search in machine learning. *arXiv*, 2015, 1502.02127. https://arxiv.org/abs/1502.02127.
(108) Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinf.* **2006**, *7* (1), 91.
(109) Iizuka, N.; Oka, M.; Yamada-Okabe, H.; Nishida, M.; Maeda, Y.; Mori, N.; Takao, T.; Tamesa, T.; Tangoku, A.; Tabuchi, H.; et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet* **2003**, *361* (9361), 923−929.
(110) Lastra-Mejias, M.; Villa-Martinez, A.; Izquierdo, M.; Aroca-Santos, R.; Cancilla, J. C.; Torrecilla, J. S. Combination of LEDs and cognitive modeling to quantify sheep cheese whey in watercourses. *Talanta* **2019**, *203*, 290−296.
(111) Nelson, N. G.; Muñoz-Carpena, R.; Phlips, E. J.; Kaplan, D.; Sucsy, P.; Hendrickson, J. Revealing Biotic and Abiotic Controls of Harmful Algal Blooms in a Shallow Subtropical Lake through Statistical Machine Learning. *Environ. Sci. Technol.* **2018**, *52* (6), 3527−3535.
(112) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N.; Baker, N. In *How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions?*, 2018 IEEE Winter Conference on Applications of Computer Vision, 2018; 2018; pp 1340−1349.

https://doi.org/10.1021/acs.est.1c01339
Environ. Sci. Technol. 2021, 55, 12741−12754

(113) Burkov, A. *The Hundred-Page Machine Learning Book*; Andriy Burkov Canada, 2019; Vol. *1*.
(114) Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, 2019.
(115) Lundberg, S. M.; Lee, S.-I. In *A Unified Approach to Interpreting Model Predictions*; Advances in Neural Information Processing Systems, 2017; pp 4765−4774.
(116) Ribeiro, M. T.; Singh, S.; Guestrin, C. In *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016; 2016; pp 1135−1144.
(117) Wang, Q.; Zhang, J.; Song, S.; Zhang, Z., Attentional neural network: Feature selection using cognitive feedback. *arXiv*, 2014, 1411.5140. https://arxiv.org/abs/1411.5140.
(118) Hu, B.; Shi, C.; Zhao, W. X.; Yu, P. S. In *Leveraging Meta-Path Based Context for Top-n Recommendation with a Neural Co-attention Model*, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018; 2018; pp 1531−1540.
(119) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **2001**, *29*, 1189−1232.
(120) Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **2012**, *6* (4), 1−21.
(121) Hand, D.; Christen, P. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing* **2018**, *28* (3), 539−547.
(122) Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, *27* (8), 861−874.
(123) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405* (2), 442−451.
(124) Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **2012**, *13* (2), 281−305.
(125) Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; Chatila, R.; Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **2020**, *58*, 82−115.
(126) Holzinger, A.; Kieseberg, P.; Weippl, E.; Tjoa, A. M. In *Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: from Machine Learning to Explainable AI*, International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 2018; Springer: Cham, 2018; pp 1−8.
(127) McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, No. 116, 11624−11629.
(128) Sutton, C.; Boley, M.; Ghiringhelli, L. M.; Rupp, M.; Vreeken, J.; Scheffler, M. Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **2020**, *11* (1), 4428.
(129) Lahrich, S.; Laghrib, F.; Farahi, A.; Bakasse, M.; Saqrane, S.; El Mhammedi, M. Review on the contamination of wastewater by COVID-19 virus: Impact and treatment. *Sci. Total Environ.* **2021**, *751*, 142325.
(130) Bogler, A.; Packman, A.; Furman, A.; Gross, A.; Kushmaro, A.; Ronen, A.; Dagot, C.; Hill, C.; Vaizel-Ohayon, D.; Morgenroth, E.; et al. Rethinking wastewater risks and monitoring in light of the COVID-19 pandemic. *Nature Sustainability* **2020**, *3*, 981.
(131) Venugopal, A.; Ganesan, H.; Sudalaimuthu Raja, S. S.; Govindasamy, V.; Arunachalam, M.; Narayanasamy, A.; Sivaprakash, P.; Rahman, P. K.S.M.; Gopalakrishnan, A. V.; Siama, Z.; Vellingiri, B. Novel wastewater surveillance strategy for early detection of coronavirus disease 2019 hotspots. *Current Opinion in Environmental Science & Health* **2020**, *17*, 8−13.
(132) Latif, S.; Usman, M.; Manzoor, S.; Iqbal, W.; Qadir, J.; Tyson, G.; Castro, I.; Razi, A.; Boulos, M. N. K.; Weller, A.; et al. Leveraging data science to combat covid-19: A comprehensive review. *IEEE Transactions on Artificial Intelligence* **2020**, *1* (1), 85−103.
(133) Bhalla, N.; Pan, Y.; Yang, Z.; Payam, A. F. Opportunities and challenges for biosensors and nanoscale analytical tools for pandemics: COVID-19. *ACS Nano* **2020**, *14* (7), 7783−7807.
(134) Mollalo, A.; Rivera, K. M.; Vahedi, B. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. *Int. J. Environ. Res. Public Health* **2020**, *17* (12), 4204.
(135) Al-Jarrah, Y. O.; Yoo, P. D.; Muhaidat, S.; Karagiannidis, G. K.; Taha, K. Efficient machine learning for big data: A review. *Big Data Research* **2015**, *2* (3), 87−93.
(136) L'heureux, A.; Grolinger, K.; Elyamany, H. F.; Capretz, M. A. Machine learning with big data: Challenges and approaches. *IEEE Access* **2017**, *5*, 7776−7797.
(137) McCullagh, P. What is a statistical model? *Annals of statistics* **2002**, *30*, 1225−1267.
(138) Noyes, P. D.; Friedman, K. P.; Browne, P.; Haselman, J. T.; Gilbert, M. E.; Hornung, M. W.; Barone, S., Jr.; Crofton, K. M.; Laws, S. C.; Stoker, T. E.; et al. Evaluating chemicals for thyroid disruption: Opportunities and challenges with in vitro testing and adverse outcome pathway approaches. *Environ. Health Perspect.* **2019**, *127* (9), 095001.
(139) Weichert, D.; Link, P.; Stoll, A.; Rüping, S.; Ihlenfeldt, S.; Wrobel, S. A review of machine learning for the optimization of production processes. *International Journal of Advanced Manufacturing Technology* **2019**, *104* (5), 1889−1902.
(140) Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer, 2013; Vol. *26*.
(141) Shorten, C.; Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **2019**, *6* (1), 60.
(142) Bühlmann, P.; Van De Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*; Springer Science & Business Media, 2011.
(143) Rexstad, E.; Innis, G. S. Model simplification—three applications. *Ecol. Modell.* **1985**, *27* (1−2), 1−13.
(144) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **2014**, *15* (1), 1929−1958.
(145) Yao, Y.; Rosasco, L.; Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation* **2007**, *26* (2), 289−315.
(146) Couzin-Frankel, J. Artificial intelligence could revolutionize medical care. But don't trust it to read your x-ray just yet. *Science* **2019**, DOI: 10.1126/science.aay4197.
(147) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533* (7601), 73−76.
(148) Charidimou, A.; Zonneveld, H. I.; Shams, S.; Kantarci, K.; Shoamanesh, A.; Hilal, S.; Yates, P. A.; Boulouis, G.; Na, H. K.; Pasi, M.; et al. APOE and cortical superficial siderosis in CAA: Meta-analysis and potential mechanisms. *Neurology* **2019**, *93* (4), e358−e371.
(149) Zhu, J.-J.; Dressel, W.; Pacion, K.; Ren, Z. J. ES&T in the 21st Century: A Data-Driven Analysis of Research Topics, Interconnections, And Trends in the Past 20 Years. *Environ. Sci. Technol.* **2021**, *55* (6), 3453−3464.
(150) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; et al. A deep learning approach to antibiotic discovery. *Cell* **2020**, *180* (4), 688−702. e13.
(151) Wills, I., The Edisonian Method: Trial and Error. In *Thomas Edison: Success and Innovation through Failure*; Springer, 2019; pp 203−222.