Environmental Science & Technology

pubs.acs.org/est

Article

Predicting Heavy Metal Adsorption on Soil with Machine Learning and Mapping Global Distribution of Soil Adsorption Capacities

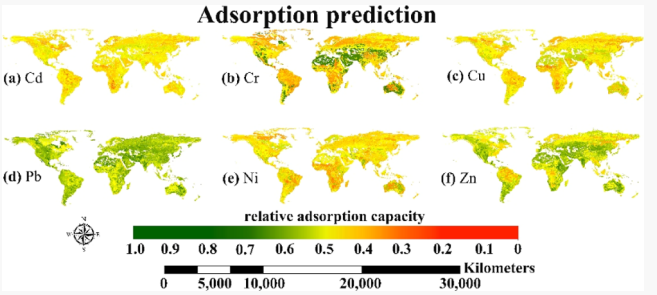
Hongrui Yang, Kuan Huang, Kai Zhang, Qin Weng, Huichun Zhang,* and Feier Wang*

Cite This: *Environ. Sci. Technol.* 2021, 55, 14316−14328

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Studying heavy metal adsorption on soil is important for understanding the fate of heavy metals and properly assessing the related environmental risks. Existing experimental methods and traditional models for quantifying adsorption, however, are time-consuming and ineffective. In this study, we developed machine learning models for the soil adsorption of six heavy metals (Cd(II), Cr(VI), Cu(II), Pb(II), Ni(II), and Zn(II)) using 4420 data points (1105 soils) extracted from 150 journal articles. After a comprehensive comparison, our results showed that the gradient boosting decision tree had the best performance for a combined model based on all the data. The Shapley additive explanation method was used to identify the feature importance and the effects of these features on the adsorption, based on which six independent models were developed for the six metals to achieve better model performance than the combined model. Using these independent models, the global distribution of heavy metal adsorption capacities on soils was predicted with known soil properties. Reversed models, including one combined model for all the six metals and six independent models, were also built using the same data sets to predict the heavy metal concentration in water when the adsorbed amount is known for a soil/sediment.



KEYWORDS: *global distribution, heavy metals, machine learning models, predictive modeling, soil adsorption capacity*

INTRODUCTION

Heavy metals in soil pose potential risks to the environment due to their inherent accumulative and non-degradable properties. Soil can also substantially affect the transport of heavy metals to water, animals, and plants[1−3] by regulating heavy metal mobility and availability, during which soil adsorption is one of the main processes.[4,5] As such, soil adsorption can reduce their environmental risks.[5,6] For example, heavy metals in soil pose low environmental risks in some regions even though the heavy metal contents exceed the standard limits.[7] Therefore, estimating the adsorption capacity of heavy metals on soil is important to assess their environmental risks and develop appropriate soil remediation strategies.[2] Moreover, the estimated adsorption capacities of different soils can help formulate strategies to mitigate the impact of heavy metals on the environment, for example, only launching the industries with high potentials of heavy metal pollution in places where the soils have high adsorption capacities. Similarly, the risk of heavy metals on agriculture can be minimized by knowing the adsorption capacities of the soil and the tolerance of crops to the heavy metals.

Soils have different properties such as pH, cation-exchange capacity (CEC), clay content, and organic carbon content (OC).[5,8−10] This heterogeneity makes the adsorption of heavy metals vary considerably.[11,12] Batch experiments are traditional ways to determine adsorption on soils, which are time-consuming and inefficient.[2,13] An alternative approach is to estimating the adsorption capacity using models.[2] Adsorption isotherms, such as the Freundlich and Langmuir equations,[14] are a traditional way to model adsorption equilibrium.[15] Many studies have used these traditional models with different combinations of independent variables (*e.g.*, soil properties) to predict the adsorption of heavy metals on soil.[16−19] However, adsorption experiments are essential to obtain the adsorption capacities prior to modeling, so the applicability domain of the isotherms is rather narrow.[10] Additionally, a combination of independent variables needs to be decided prior to developing desirable models for specific types of soils because of soil heterogeneity. This also makes it difficult to use these traditional models to quantify the adsorption characteristics of soils on a large scale. As a mechanistic model, surface complexation models (SCMs) have been widely employed to simulate the adsorption of different adsorbates onto different adsorbents.[20−22] SCMs can provide information about surface complexation reactions with a set of equilibrium constants

Received: April 15, 2021
Revised: September 20, 2021
Accepted: September 21, 2021
Published: October 7, 2021

insensitive to the change in solution conditions, which is important for robust prediction.[20−22] Many studies have successfully used SCMs to model the interactions of heavy metals with soils or soil constituents.[23] However, predictions using SCMs are based on three assumptions—adsorption occurs on the surfaces that have ligand functional groups; adsorption reactions follow the law of mass conservation; and surface coordination reaction determines the surface charge distribution.[24] These assumptions might lead to poor performance of SCMs in complex systems and thus limit the applicability of SCMs. Indeed, studies on SCMs mainly focus on materials with relatively well-defined surface groups.[24] Therefore, to better model the adsorption of different adsorbates to complex adsorbents such as soil, new approaches should be explored to overcome the above limitations.

As a powerful tool for uncovering hidden relationships, machine learning approaches have been increasingly applied to study environmental problems because of their low cost, high prediction accuracy, and robustness.[13,25−31] Traditional learning models, such as classification and regression trees (CART), linear regression (LR), stochastic gradient descent regressor (SGDRegressor), support vector regression (SVR), ridge regression (Ridge), and K-nearest neighbors (KNN), have been widely used in many fields.[32−34] These traditional models are developed with single algorithms and have relatively low prediction precision in some studies.[25] Ensemble models consisting of multiple learning algorithms—such as regression tree-based ensemble learning models, including extremely randomized trees (ET), random forest (RF), gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost)—may be employed sometimes to improve accuracy and obtain better predictive performance.[25,35−38]

Despite the extensive applications of machine learning in the field of environmental research, very limited work has been conducted on the adsorption of heavy metals on soil. Bazoobandi *et al.* adopted an artificial neural network (ANN) to predict the contents of Cd and Pb in soil and identified OC as the most significant predictor.[39] Tan *et al.* employed RF to develop a hyperspectral estimation model to accurately predict the spatial distribution of Cr, Cu, and Pb in agricultural soils.[40] Jia *et al.* combined RF with the fuzzy *k*-means method to predict the concentrations of Cr, Pb, Hg, and As in soils and then partitioned the study area into four subregions with different potential risk levels based on the predicted heavy metal concentrations.[41] For the prediction of adsorption on soils, however, to the best of our knowledge, only one study has reported such an effort using machine learning, that is, a study by Anagu *et al.* using ANN to predict the adsorption of nine heavy metals based on the data collected from 133 agricultural sites in Germany.[2] However, the used data volume (133 data points) is small; the soil diversity (69 soil types) is not high; and the performance of the obtained model is relatively poor for one of the metals (Cr: R^2 = 0.79), which limit the applicability of the model. To overcome these limitations, a much larger amount of data covering much more diverse types of soils should be collected to make the models more robust and widely applicable.

The aims of this study are to: (i) build a comprehensive data set for the adsorption of six heavy metals [*i.e.*, cadmium (Cd(II)), chromium (Cr(VI)), copper (Cu(II)), lead (Pb(II)), nickel (Ni(II)), and zinc (Zn(II))] on different soils by collecting a total of 4420 data points (1105 soil types) from 150 relevant references; (ii) based on the above data set, investigate the performance of six traditional learning models (*i.e.*, CART, LR, SGDRegressor, SVR, KNN, and Ridge) and four ensemble models (*i.e.*, ET, RF, GBDT, and XGBoost) in the prediction of heavy metal adsorption on soils; (iii) using the best performing models, identify the key factors from the properties of soil, adsorption systems, and heavy metals that are important for the adsorption; and (iv) predict the relative adsorption capacities of the six heavy metals on a global scale for the first time. These findings can help formulate strategies for soil remediation, regional land-use planning, and risk assessment. An online predictor with a graphical user interface is available in the adsorption section of ChemAI at https://www.chemai.aropha.com/.

MATERIALS AND METHODS

Literature Search Protocol. A comprehensive literature search was conducted using the Web of Science to obtain data on the adsorption of heavy metals onto different types of soils, using the following search terms, where "TS" represents the article theme:

TS = [(soil OR soils) AND (adsorption OR sorption) AND (metal OR metals OR Cd OR Cr OR Cu OR Pb OR Ni OR Zn OR cadmium OR chromium OR copper OR lead OR nickel OR zinc)].

Study Selection. The searched relevant studies were ranked by relevance and screened by examining the sections of materials and methods and results and discussion to determine the suitability of the obtained results. A total of 150 studies (Table S1, Supporting Information) were selected based on the following criteria: (1) soil properties including pH, CEC, OC, and clay content were measured and reported; (2) an electrolyte was used to maintain the ionic strength in the batch sorption experiments, and the ionic strength of all other substances accounted for less than 5% of the total ionic strength; (3) adsorption system properties, including solution pH, the equilibrium concentration of heavy metal(s), solution temperature, and soil-to-solution ratio, were reported; and (4) the adsorption data of heavy metals on soil were accessible (*e.g.*, in a table or figure with exact coordinates).

Data Extraction. The collected data included the study metadata (year published, first author's last name); soil properties including soil pH, CEC, OC, and clay content; adsorption system properties including solution pH, the concentrations of background electrolytes, heavy metal concentrations at equilibrium, solution temperature, and soil-to-solution ratio; and the adsorption data of the heavy metals on soil. Data in tables and texts were extracted by transcription, while the data presented graphically were extracted manually using PlotDigitizer (http://plotdigitizer.sourceforge.net/). From the 150 studies, 4420 adsorption data points associated with 1105 soil types and six heavy metals (*i.e.*, Cd, Cr, Cu, Pb, Ni, and Zn) were mined. There are 1092, 436, 764, 888, 512, and 728 data points for Cd, Cr, Cu, Pb, Ni, and Zn, respectively.

10 Machine Learning Models. To identify the best model for the prediction of the adsorption of the six heavy metals on soil, six traditional and four ensemble models were selected and compared for the prediction performance. The detailed information about these 10 machine learning models can be found in Texts S1 and S2.

Model Development. The procedure of the model development included: (1) unifying the units of each variable;

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

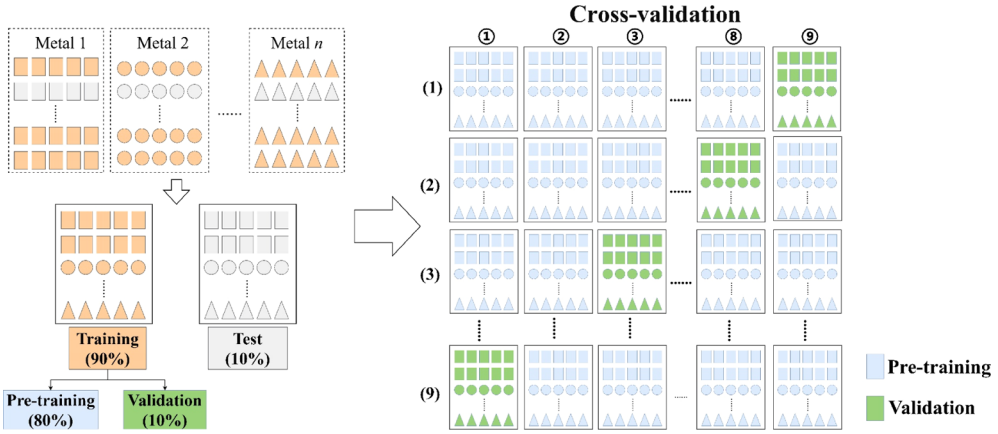


Figure 1. Left: Data-splitting approach to minimizing data leakage and achieving good model performance. For each metal, each row represents one group of data (*i.e.*, one isotherm) containing several data points; these groups were split into a training data set (orange color) and a test data set (gray color) in a ratio of 9:1. The training and test data sets for different metals (*i.e.*, different geometrical shapes) were combined into the final training and test data sets, respectively. Right: The 9-fold cross validation was conducted on the above training data set, during which the training data set was further split into pre-training (blue color) and validation (green color) data sets in a ratio of 8:1.

(2) transforming the output adsorption data into the natural logarithm form (Ln-adsorption); (3) splitting the data into training and test sets in the ratio of 9:1; (4) tuning the parameters of the 10 algorithms and finding the best parameters for each; and (5) quantifying the model performance. To reduce the over-fitting risk,[25] 9-fold cross validation (details given in Text S3) was conducted on the training data set, where the training set was further split into pre-training and validation sets in the ratio of 8:1.[42] During the modeling, the input data included (i) four descriptors for soil properties, namely, pH of soil, CEC (cmol/kg), OC (%), and clay content (%), which are commonly used soil properties in the selected studies; (ii) three descriptors for the heavy metals, namely, the first ionization energy (IE, kJ/mol), ionic radius (radius, Å), and hydrated ionic radius (hydra_radius, Å),[43] which were selected because they are among important properties of heavy metals and did not correlate with each other (details given in Text S4 and Table S2); and (iii) five descriptors for the adsorption system, namely, the equilibrium concentration (C_e, mg/L), solution pH, ionic strength (I, mol/L), temperature (T, °C), and soil-to-solution ratio (g/mL), which are key factors in determining the adsorbed amounts of adsorbates.[13,44−46] Note that the soil pH is the pH measured for soils upon collection by mixing with water. The output was the natural logarithm of the corresponding adsorbed heavy metal amount on soil (Ln-mg/g).

When developing a machine learning model, data splitting is a critical step; appropriate approaches are necessary to minimize possible data leakage and achieve good model performance. Data leakage is when information from outside the training data set is used to train the model, which makes the model learn irrelevant information and in turn invalidates the prediction performance of the model.[47] Some data points in our study were extracted from the same adsorption isotherms, which might lead to data leakage if these data points are split into pre-training, validation, and test data sets.[13] To minimize the potential data leakage, data points from the same adsorption isotherm were integrated as one group, and data splitting was conducted on these groups prior to 9-fold cross validation (Figure 1). To achieve good model performance, we then split the groups into pre-training, validation, and test data sets (Figure 1).

The model parameters were tuned with 9-fold cross validation, and the optimal configuration of the models was determined using the grid search method (Text S5). The performance of models with different configurations were evaluated by comparing R^2 values and four regression loss functions (details given in Text S6) on the validation data set, and the optimal model configuration was the one with the highest R^2 and the smallest loss function values on the validation data set.

Model Selection. To measure the deviation of our model prediction from the ground truth, we employed R^2 and four regression loss functions to evaluate the performance of 10 models and selected the model with the highest R^2 and the smallest loss function values as the final prediction model for the entire data set—referred to as the combined model hereafter. The loss functions used in our study included root-mean-square error (rmse), mean absolute error (MAE), Huber loss, and Log-cosh loss (details given in Text S6).

Identification of Key Parameters. To evaluate the importance of different descriptors on the adsorption of heavy metals on soils, the Shapley additive explanation (SHAP) method (details given in Text S7) was employed to calculate the Shapley values for each descriptor. The Shapley value is a concept in the cooperative game theory, which fairly assigns a unique distribution among the descriptors of a total surplus (*e.g.*, the predicted adsorption in this study) generated by the coalition of all the descriptors.[48] SHAP is an additive feature attribution method based on the theoretically optimal Shapley values to explain individual predictions,[49] where an individual prediction refers to one data point (*i.e.*, the adsorption of one heavy metal on one soil in a given solution system). For each data point, each descriptor has one SHAP value, which represents the effect of that descriptor on the behavior of that data point. Then, all the SHAP values for each descriptor are averaged as the mean absolute Shapley (MAS) to quantify the overall impact of the descriptor; the larger the MAS value, the more significant the descriptor is in influencing the adsorption.

Independent Models for Each Metal. The final combined model can be used to predict the soil adsorption of the six heavy metals. However, the combined model might not have good prediction performance for some metals.

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

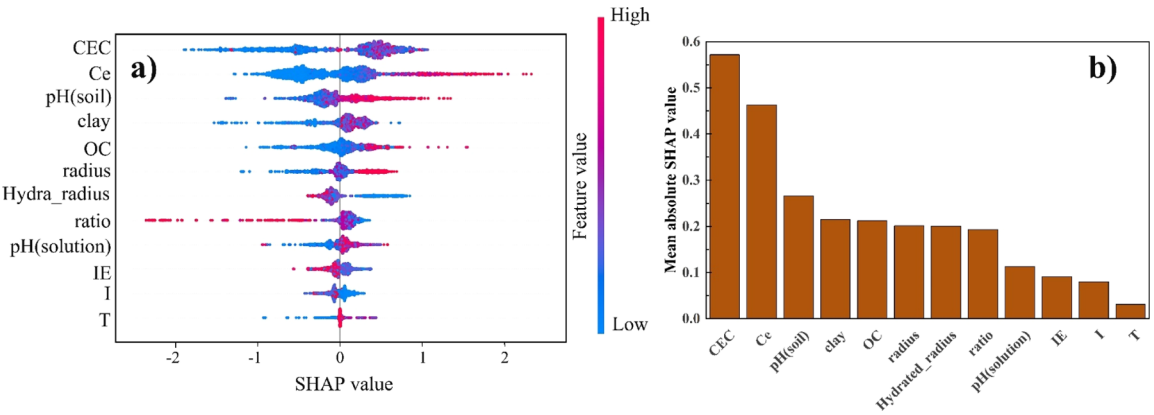


Figure 2. (a) Shapley additive explanations (SHAP) values for all the input descriptors with all the data points included, and (b) MAS values for all the descriptors in the combined model. Ratio, IE, *I*, and *T* denote the soil-to-solution ratio, the first IE, ionic strength, and temperature, respectively.

Considering significantly different adsorption mechanisms among different heavy metals onto soils, independent models for each metal were developed following the same approach. During the independent model development, the three descriptors for the heavy metals (*i.e.*, IE, radius, and hydra_radius) and the soil/solution descriptors with less influence on adsorption—based on the identified key parameters for the combined model—were eliminated from the model inputs. The model performance of independent models and the combined model were compared to determine the final prediction model for each heavy metal.

Note that we also developed 15 paired models by combining two heavy metals into one model, that is, selecting two of the six metals, which had 15 combinations. However, these paired models had less satisfactory model performance than the six metal-specific models so they are not further discussed.

Model Application—Global Spatial Distribution of Soil Adsorption Capacities. The final prediction models for each metal were used to predict the adsorption capacity of heavy metals on soil on a global scale. The values of soil properties (*i.e.*, pH, CEC, OC, and clay content) were extracted from the Harmonized World Soil Database,[50] a 30 arc-second raster database in the AcrGIS software with over 15,000 different soil mapping units that combine existing regional and national updates of soil information worldwide. The soil mapping units are grids that contain soil properties; a total of over 15,000 grids are able to represent the soil distribution around the world. To ensure that the predicted adsorbed amounts can be compared for different soils, the descriptors for the adsorption systems used in the final models were set at the same values for all samples (details given in the results and discussion section), and the solution pH was set at the same value as the soil pH. In addition, as the predicted adsorption capacity (mg/g) depends on the adsorption system, the predicted values were rescaled to 0−1 (eq 1), representing the relative adsorption capacities

$$Q_i = \frac{Q_{i,output} - Q_{min}}{Q_{max} - Q_{min}} \quad (1)$$

where Q_i is the relative adsorption capacity of soil i; $Q_{i,output}$, Q_{min}, and Q_{max} (mg/g) are the predicted adsorption of soil i and the minimum and the maximum values of all the predictions, respectively. To evaluate the potential impact of the selected descriptor values on the obtained Q_i values, soil adsorption was predicted at several values for each descriptor of the adsorption system (details given in the results and discussion section). The Q_i values at different descriptor values were then compared using the cosine similarity metric to evaluate their similarity (details given in Text S8). The higher the similarity among the Q_i values at different descriptor values, the less of impact these descriptor values have. After that, we obtained the global spatial distribution of the relative adsorption capacities for the heavy metals.

To explore the overall adsorption capacities of heavy metals in different countries, the mean relative adsorption capacity for each heavy metal is defined using eq 2

$$Q_{mean} = \frac{\sum_{i=1}^{n} Q_i \times S_i}{S} \quad (2)$$

where Q_{mean} is the mean relative adsorption capacity for a country; Q_i is the relative adsorption capacity of soil i in the country; S_i is the area of soil i; and S is the total area of the country. With such mean values, comparative analysis was conducted among countries to find those with the highest and lowest adsorption capacities, respectively. This information is significant to recognize the overall situation of countries on a global scale. The current heavy metal pollution status of some regions in Europe, Africa, and China was also analyzed and compared with their corresponding modeled adsorption capacities to evaluate the potential risks associated with these heavy metals.

RESULTS AND DISCUSSION

Combined Model Development and Comparison. After reviewing the 4420 data points extracted from 150 studies, we observed the ranges of pH (soil), CEC, OC, clay content, C_e, pH (solution), ionic strength, temperature (T), soil-to-solution ratio, and adsorbed amount being 3.07−9.70, 0.100−117.0 cmol/kg, 0.0098−63.0%, 0.4−93.1%, 0.0002−2297.9 mg/L, 2.0−12.0, 0.001−0.50 mol/L, 15−45 °C, 0.0002−0.667 g/mL, and 0.001−944.7 mg/g, respectively.

With the data-splitting approach mentioned above, the model parameters were tuned to improve the model performance and obtain the optimized models (Table S3). According to the R^2 and four loss function values of the 10 models (Table S4), the optimized GBDT model achieved the best performance and was thus selected for future discussion. Comparison between the reported and predicted values

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

indicated high reliability and robustness of the GBDT model, with the R^2 of 0.780 and 0.778, rmse of 0.913 and 0.826, MAE of 0.624 and 0.583, Huber loss of 0.0577 and 0.0536, and Log-cosh loss of 0.2810 and 0.02471 for the validation and test data sets, respectively. However, the GBDT model which was based on the entire data set (referred to as the "combined" model hereafter) had different prediction accuracy for different heavy metals (Table S5), suggesting the necessity of developing metal-specific prediction models for each metal (details given in the independent models development section).

Influence of Input Parameters on the Adsorption. In addition to improving the model performance, it is essential to interpreting the feature importance to see whether it agrees with the known mechanisms so that we can trust the model. To achieve this, the SHAP method was used.

The SHAP values of the 12 input descriptors were obtained for all the data points (Figure 2a). As shown, the horizontal position of each point is based on its SHAP value, and the color indicates the feature value of the descriptor—red and blue represent large and small values of the descriptor, respectively. For each descriptor, the SHAP values of all samples varied within a certain range. Some samples with similar feature values (*i.e.*, similar colors) have different SHAP values (*i.e.*, different horizontal positions), demonstrating that the contribution of a descriptor is not only determined by its own feature value but also strongly influenced by other descriptors. For CEC, C_e, pH (soil), clay content, OC, radius, pH (solution), and temperature, the points with low feature values (in blue) are mainly on the left side, while the points with high feature values (in red) are mainly on the right side, suggesting positive correlations between these descriptors and the predicted adsorption—higher feature values favor the adsorption. On the contrary, the points with high values (in red) for hydra_radius, ratio, the first IE, or ionic strength generally are distributed on the left side, showing their negative correlations with the predicted adsorption. These results are also consistent with the Pearson correlation coefficients between each descriptor and the predicted adsorption (Table S6).

The MAS value was then calculated for each input descriptor based on all the data points to quantify the overall importance of each descriptor, which was found to follow the order of CEC > C_e > pH (soil) > clay > OC > radius > hydra_radius > ratio > pH (solution) > the first IE > ionic strength > temperature (Figure 2b).

The above results showed that all the four soil properties—CEC, pH, OC, and clay content—are among the most important features in deciding the output (*i.e.*, soil adsorption of heavy metals). Soil CEC, a measure of the amount of total exchangeable cations and the total negative surface charge of soil,[51] showed considerable influences on the predicted adsorption according to the SHAP analysis (Figure 2). Indeed, extensive studies have reported positive correlations between CEC and the adsorption of heavy metals on soil.[10,51−53] Soil pH is another important factor, in agreement with many experimental studies.[8,54] At low pH, H^+ in soil can compete strongly with metal ions for active adsorption sites leading to attenuated adsorption.[55,56] Soil pH also affects the charge status of soil surfaces; higher soil pH usually leads to more negatively charged sites and thus better adsorption. Clay minerals tend to have small particle sizes with very high specific surface areas and have the ability to sequester heavy metals through complexation reactions and electrostatic attraction.[9,57] Soil OC offers functional groups to complex with metal ions,[58] which can significantly enhance the adsorption capacity. McBride *et al.* observed an increase in the CEC and therefore the number of adsorption sites upon the amendment of soils with OC.[52] In addition, neutral metal hydroxide species are more hydrophobic than charged metal species and, hence, may prefer to be associated with hydrophobic OC on soil surfaces.[59]

For the effect of solution properties on adsorption, C_e is the most influential factor. The positive correlation between C_e and adsorption suggests that the adsorption of metals on soil would increase with increasing C_e. This agrees with many previous studies.[18,60,61] For example, after developing an ANN model to predict the adsorption of metals onto soils and conducting sensitivity analysis to recognize the relative importance of each input descriptor, Anagu *et al.* showed that C_e is the most important variable.[2]

The effect of solution pH on the adsorption of heavy metals on soil surfaces can be understood in terms of two common adsorption mechanisms:[23] specific inner-sphere complexation between heavy metals and surface functional groups and non-specific outer-sphere electrostatic interactions. Soil surfaces usually contain multiple functional groups, with the most common ones being silanol, carboxyl, carbonyl, phenolic, and inorganic hydroxyl groups.[23] Some minerals (*e.g.*, aluminosilicates) may also contain exchangeable ion-bearing sites in addition to protons. These interactions are highly dependent on the solution pH, as supported by the SHAP analysis results which indicate that higher solution pH is beneficial for heavy metal adsorption, in agreement with other experimental studies.[23] This is due to the fact that an increase in pH generally increases the number of negatively charged surface sites, which can facilitate the complexation and ion exchange between the soil surface and heavy metals. Higher pH can also lead to increased hydroxylation of the metal species, the adsorption of which can effectively contribute to the overall adsorption of some heavy metals, such as Pb and Cu.[62] In addition, heavy metals tend to be in the form of free ions under typical soil pH conditions (Figure S1), which substantially inhibits their electrostatic interactions with positively charged soil surfaces at lower pH. As different heavy metals have different pK_a values, their speciation changes respond differently to pH changes. For example, Cd^{2+} and Cu^{2+} dominate the system when the pH is less than 10 and 6.7, respectively (Figure S1). This may also change when other anions/ligands (*e.g.*, Cl^-) complex with them. Therefore, the effect of solution pH on heavy metal adsorption on soil is largely dependent on the types of soils, heavy metals, or co-existing anions.

Numerous studies have reported that the soil-to-solution ratio can significantly influence soil adsorption of heavy metals, with a higher soil-to-solution ratio leading to a decrease in the adsorption, which is consistent with our SHAP analysis result.[46,63,64] The higher the soil-to-solution ratio, the more soil exists in the solution, resulting in less adsorption per unit mass of soil (mg/g), despite the higher total adsorbed amount (mg). However, this influence is subject to soil properties.[63,65,66] Many studies demonstrated that the heavy metal distribution coefficient (*i.e.*, the ratio of metal adsorbed on soil to that in solution at equilibrium) is less sensitive to the soil-to-solution ratio at low soil pH, presumably because of the low adsorption capacity at low pH.[64] Besides, the soil-to-solution ratio has an effect on the concentration of dissolved organic

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

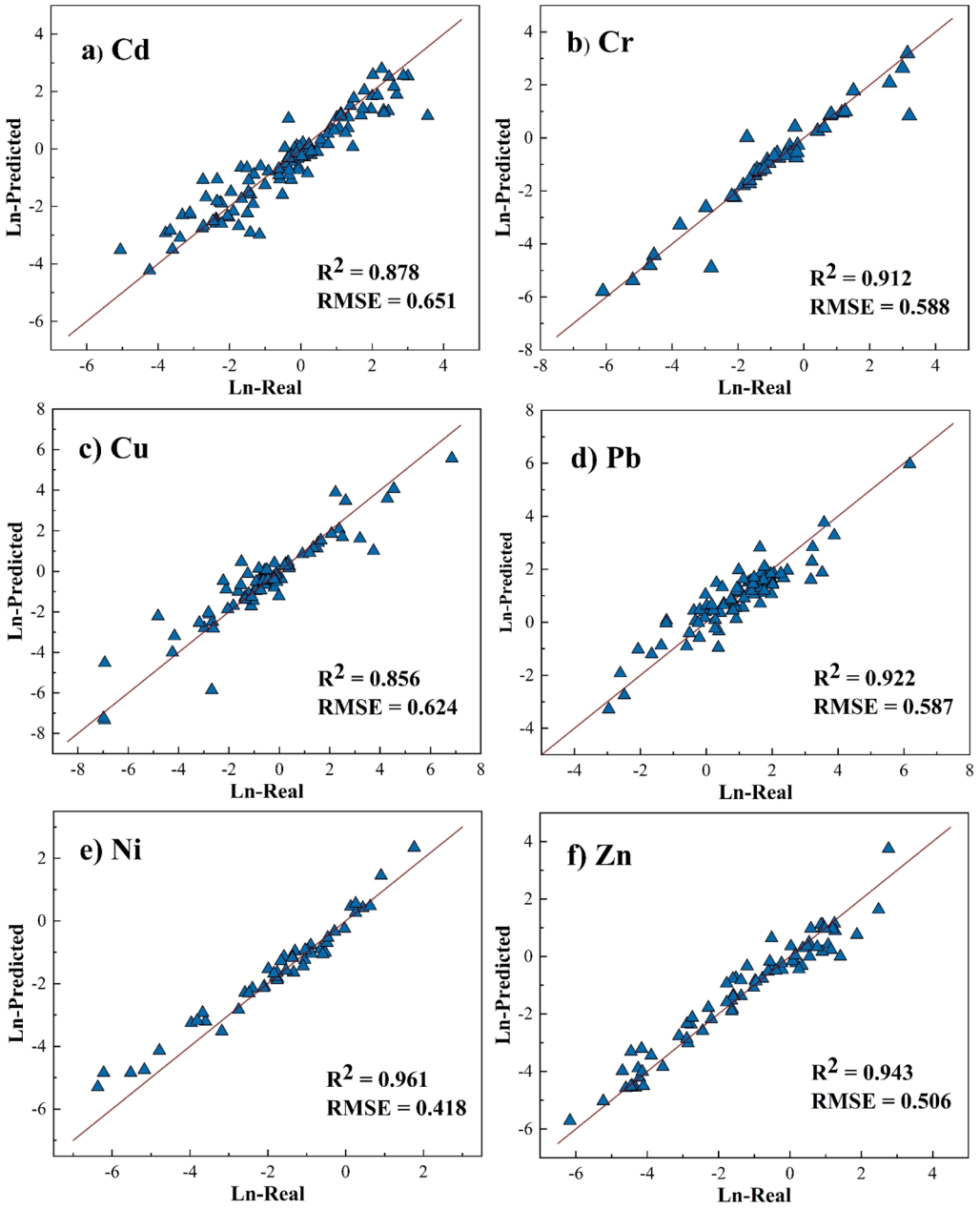


Figure 3. Performance of the (a) independent-Cd model; (b) independent-Cr model; (c) independent-Cu model; (d) independent-Pb model; (e) independent-Ni model; and (f) independent-Zn model in predicting the soil adsorption of the respective metal (Ln-Predicted) for the corresponding test data set (Ln-Real). For the model output, the soil adsorbed amount (mg/g) was transformed into natural logarithm (ln-mg/g).

carbon, which may compete or complex with heavy metals to influence the soil adsorption to heavy metals.[65,66]

Solution ionic strength and solution temperature may not be as important as C_e, soil-to-solution ratio, or solution pH, but they can also affect the adsorption. Studies have widely reported that at certain pH, higher ionic strength is beneficial for the adsorption, while at other pH conditions, this effect may be reversed.[67−70] Such a pH may be near the pHzpc (pH of zero point of charge) of the soil.[67] For example, at pH greater than pHzpc, increasing ionic strength decreased Cd adsorption regardless of the soil type because the cations competed with Cd for the active adsorption sites.[71] However, at pH higher than a certain level, increasing ionic strength increased the adsorption of As(V), and *vice versa*.[70] This is because the number of cations in the adsorption plane increases with increasing ionic strength, resulting in a less negative potential at the adsorption plane and thus facilitating the adsorption of As(V) by the soil.[70] Therefore, the ionic strength effect is largely dependent on the types of soils and heavy metals. Solution temperature had a marginally positive influence on soil adsorption according to our SHAP analysis. Solution temperature influences soil porosity; increasing temperature causes soil to swell, which results in the better penetration of heavy metals into the pores.[45] Moreover, soil adsorption depends on the interactions between the heavy metal and functional groups of organic matter in the soil. The interactions tend to strength at higher solution temperature, resulting in improved soil adsorption.[45]

As for the heavy metal properties, a larger hydrated ionic radius or radius of heavy metals is reported to decrease the

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

Table 1. Descriptive Statistics of the Data Ranges of Four Soil Properties for the Six Model Data Sets and for Worldwide Soils

	data range[a]				exceeding percentage (%)[b]				
	pH	CEC	OC	clay content	pH	CEC	OC	clay content	total[c]
Cd	3.1−9.7	0.1−117.0	0.027−14.6	0.5−93.1	0.080	0.337	1.549	0.006	1.623
Cr	3.6−8.7	1.5−97.7	0.048−16.8	0.5−69.0	0.698	0.784	1.580	0.453	1.929
Cu	3.3−8.6	0.1−117.0	0.010−63.0	0.5−93.1	0.980	0.337	0.000	0.006	1.274
Pb	3.3−8.9	0.1−97.7	0.010−16.8	0.4−69.0	0.606	0.343	1.513	0.453	2.119
Ni	4.0−8.6	0.1−117.0	0.058−16.8	0.5−82.0	1.298	0.337	1.598	0.037	2.860
Zn	3.3−8.7	0.1−117.0	0.034−16.8	0.5−70.0	0.606	0.337	1.525	0.416	2.548
World	3.0−10.6	1.0−134	0.01−47.2	1.0−94.0					

[a]The data ranges of the four soil properties in the collected data sets for the six heavy metals and for the soil samples in the Harmonized World Soil Database (shown as "world"). [b]The data ranges of the four soil properties in the Harmonized World Soil Database (shown as "world") were compared with those of the individual collected data sets. Exceeding percentage (%) = ×100%. [c]The percentage of soil mapping units whose values for any of the soil properties exceed the data range of pH, CEC, OC, or clay content in the individual data sets.

adsorption capacity.[72−74] This is because a large hydrated ionic radius can reduce the binding strength of metal−OC or metal−clay complexation, lower the charge density to reduce the electrostatic interactions between the heavy metals and soil surfaces, and/or limit the access of heavy metals to small pores. A high first IE would reduce atom ionization, which is one of the significant steps in adsorption through the formation of metal−organic complexes.[75] Therefore, a high first IE would lead to low adsorption of heavy metals on soil. This is similar to how pH affects the charges and speciation of heavy metals.[76]

Independent Model Development. The combined model for the six heavy metals had poor prediction accuracies for some metals, especially for Cd and Cr (details given in the combined model development and comparison section), so it is necessary to develop independent models for each metal to see whether better accuracy can be obtained. Based on the feature importance of the 12 inputs on the soil adsorption from the SHAP analysis (Figure 2), the two inputs with the least importance (*i.e.*, ionic strength and solution temperature) and the three heavy metal properties (*i.e.*, radius, hydra_radius, and the first ionic IE) were eliminated; seven descriptors including soil pH, CEC, OC, clay content, C_e, solution pH, and soil-to-solution ratio were employed as the final model inputs (descriptive statistics given in Table S7). Considering the good performance of ET, GBDT, RF, and XGBoost compared with the other six learning methods (Table S4), these four algorithms were used to develop independent models for each metal, following the same procedure as in the combined model development section. Based on the model performance of the four models with ET, GBDT, RF, or XGBoost for each metal (Table S8), the GBDT models had the best performance and were thus selected as the final independent models for all metals.

Compared with the combined model, the independent models had better performance (Table S8 and Figure 3). This is because the predictions of the combined model were a global optimal solution rather than a local optimal solution, that is, the combined model had an overall good prediction accuracy but failed to obtain the most accurate predictions for some metals. Besides, most global optimization methods cannot guarantee a global minimum, especially in high-dimension problems.[77,78] On the contrary, the six independent models captured the relationship between several key features and the soil adsorption for each metal, which could provide much more confidence of learning for each metal.[79] In addition, the independent models had fewer model inputs because the eliminated features were the least important, which was beneficial for reducing the dimensionality and improving the model performance.[80] It is also likely that the selected three features for the heavy metals were not able to reflect their different adsorption mechanisms, which is understandable giving the complexity in their adsorption mechanisms.[63,64] As a result, individual models for each metal achieved better predictive performance. To validate these independent models, the SHAP analysis was conducted, and the results agreed well with the known mechanisms (details given in Text S9 and Figure S2).

Global Spatial Distribution of Relative Adsorption Capacities. Soils around the world are highly heterogeneous, which can lead to substantial variations in heavy metal adsorption.[11] The adsorption capacity of heavy metals is a significant factor in optimizing distribution of industries in order to mitigate the impact of pollutants on the environment—for example, only discharging heavy metals into soils with high adsorption capacities, or farming based on the adsorption capacities of the soil and the tolerance of crops to heavy metals. To evaluate the global spatial distribution of soil adsorption capacities, the six independent models were used to predict the soil adsorption capacities of the six metals. To verify the applicability of the six models on the global scale, comparative analysis was conducted to compare the data ranges of the soil properties between the collected data sets and worldwide soils. The results (Table 1) showed that the worldwide soil properties had slightly wider data ranges than those in the collected data sets. The distributions in the box plots (Figure S3) indicated that the data ranges of the soil properties used in the models fell within 5−95% of the data ranges of the world soil properties. Moreover, in the Harmonized World Soil Database[50] (details given in the model application—global spatial distribution of soil adsorption capacities section), only a small portion of the soil mapping units had soil properties outside the data ranges of the collected data sets (Table 1). The soil properties of 96.26% of the mapping units were fully covered by the six collected data sets. Besides, the sampling sites of the 150 cited studies distributed all over the world, including more than 40 countries and six continents (except for Antarctica, details given in Table S1), suggesting the global representativeness of the data. Therefore, we believe that it is valid and reliable to apply the six independent models to the prediction of the global distributions of soil adsorption.

In the process of model applications, worldwide soil properties were collected and then imported into the six independent models for prediction. As mentioned in the model

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

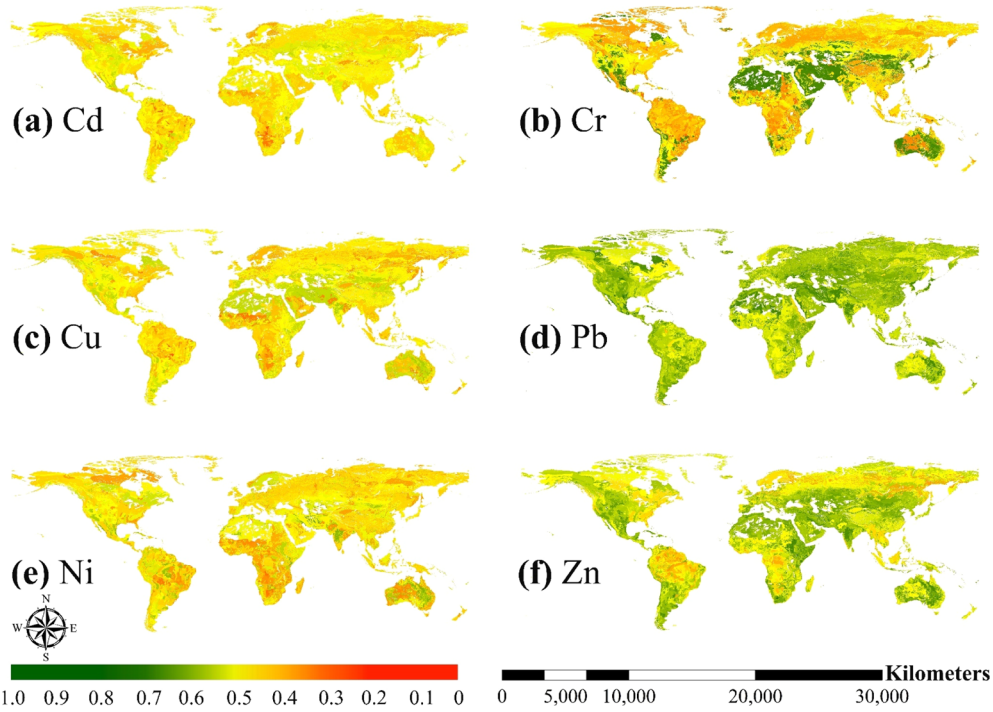


Figure 4. Global spatial distribution of the relative adsorption capacities of (a) Cd, (b) Cr, (c) Cu, (d) Pb, (e) Ni, and (f) Zn on worldwide soils. Higher adsorption capacities generally mean safer environments, so we selected greener color for such soils. On the contrary, warmer color is used for soils with lower adsorption capacities.

Table 2. Mean Values of pH, OC, CEC, and Clay Content for Soils with the Lowest or Highest Relative Adsorption Capacities[a]

	bottom 20% soil samples				top 20% soil samples			
metal	pH (soil)	OC (%)	CEC (cmol/kg)	clay (%)	pH (soil)	OC (%)	CEC (cmol/kg)	clay (%)
Cd	5.97	1.19	9.64	11.41	6.92	2.84	25.93	35.56
Cr	5.57	0.87	16.06	24.41	7.88	3.11	16.52	26.50
Cu	5.65	0.73	5.74	10.45	7.85	1.04	17.95	27.62
Pb	6.45	0.68	9.54	15.79	7.18	3.72	27.28	34.69
Ni	5.77	0.66	5.36	15.89	7.06	2.39	22.54	32.90
Zn	5.01	0.7	12.16	15.94	7.47	3.10	20.80	34.46

[a]The bottom 20% soil samples represent the soil samples with the lowest relative adsorption capacities, and the top 20% soil samples represent the soil samples with the highest relative adsorption capacities.

application—global spatial distribution of soil adsorption capacities section, the C_e and soil-to-solution ratio were set at the same values for all the input samples, and the solution pH was set at the same value as the soil pH. When evaluating the impact of the selected C_e or soil-to-solution ratio values on the model outputs, we observed high similarity among the rescaled prediction values under different C_e (including 0.1, 0.5, 1, 10, 20, 50, and 100 mg/L) or soil-to-solution ratio values (including 0.01, 0.05, 0.1, 0.2, 0.5, and 1 g/mL) (Tables S9 and S10), suggesting that the selected C_e or soil-to-solution ratio levels had little effect on the model-predicted relative adsorption capacities. We then fixed the C_e (0.1 mg/L) and soil-to-solution ratio (0.1 g/mL) and calculated the global spatial distribution of the relative adsorption capacities of worldwide soils from the six independent models.

As shown in Figure 4, the overall relative adsorption capacity of Pb is slightly higher than that of Zn and much higher than those of four other metals. Similarly, Elbana *et al.* quantified the adsorption of five heavy metals (*i.e.*, Cd, Cu, Ni, Pb, and Zn) in 10 soils with batch adsorption experiments, and the results indicated that Pb had the strongest adsorption in all the soils.[10] Indeed, Pb is often regarded as immobile for its high sorption onto most soils.[81] In contrast, the overall relative adsorption capacities of Cd, Cu, and Ni were lower than those of the other metals, which agrees with Hou *et al.*'s report that Cd had a comparatively low adsorption potential.[82] Moreover, the soil relative adsorption capacity for Cr varied greatly on the global scale compared with those of other heavy metals.

Overall, soils with low relative adsorption capacities are primarily in the northern North America, the northern South America and central Africa, while those in northern Africa and southern Asia generally have much higher adsorption capacities. To examine the differences between soils with low and high relative adsorption capacities, all the soil samples were sorted by their relative adsorption capacities. By comparing the mean values of the four soil properties for the bottom 20% soil samples (*i.e.*, soils with the lowest relative adsorption capacities) and for the top 20% soil samples (*i.e.*, soils with the highest relative adsorption capacities), the results showed that the soils with high adsorption capacities have

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

higher pH, OC, CEC, and clay contents than the soils with low adsorption capacities (Table 2), which is also consistent with the SHAP analysis results above.

To further investigate the relative adsorption capacities of soils in different countries, the mean relative adsorption capacities of each metal were calculated for different countries based on eq 2. As shown in Figure S4, the mean values decreased in the order of Pb (0.673) > Zn (0.615) > Cr (0.497) > Cu (0.467) > Cd (0.465) > Ni (0.448), which is consistent with the trend in the worldwide soils without setting country boundaries. The coefficient of variations—the ratio of the standard deviation to the mean—decreased in the order of Cr (0.340) > Ni (0.163) > Zn (0.153) > Cu (0.118) > Cd (0.132) > Pb (0.122). These results showed that the mean value of Pb maintained at the highest level but with the smallest variation, demonstrating the highest relative adsorption capacity of Pb on the global scale with the smallest variation from country to country. The coefficient of variation of the soil relative adsorption capacity for Cr was substantially higher than those of other metals. Overall, knowing the soil relative adsorption capacities can help countries set their national quality standards for heavy metals in soil (details given in Text S10 and Figure S5).

Combined Look at Heavy Metal Pollution and Soil Adsorption around the World. The environmental risks associated with heavy metals are determined by both the heavy metal contents and the adsorption capacities of the soil.[5] Therefore, it is necessary to consider soil adsorption capacities to quantify bioavailable heavy metals for accurate environmental risk assessment.

About 50% of contaminated sites globally are contaminated by heavy metals,[83] and the majority of these sites are in developed countries due to the extensive applications of heavy metals in industrial processes.[84,85] According to the European Environmental Agency (EEA), there are about 250,000 heavy metal-polluted sites in the EEA member countries, and approximately 0.5 million sites in Europe are highly contaminated.[83] However, the assessment was solely based on the heavy metal levels in soil and ignored the soil adsorption capacities, which may give inaccurate results. To this end, both heavy metal contents and the adsorption capacities of the soil should be considered. According to the reported heavy metal contents in soils in Europe, the North-eastern and Eastern-Central Europe suffer less contamination from heavy metals, while most areas in Western-Europe and the Mediterranean have concentrations exceeding the set threshold for at least one heavy metal.[86] Based on the spatial distribution of the relative soil adsorption capacities in Europe (Figure 4), soils in North-eastern and Eastern-Central Europe have much higher adsorption capacities for all the heavy metals than soils in Western-Europe. Therefore, the environmental risks might be much lower in the former regions. As for the Mediterranean, although this region has high concentrations of soil heavy metals, the environmental risks might be lower than expected due to the high adsorption capacities.

Cd, with much lower soil-relative adsorption capacities, is undetectable in most soil samples in Europe,[86] which indicates that the overall environmental risks caused by Cd might be low. However, located in Western-Europe, Ireland has the highest mean Cd concentration.[86] Considering the lower soil adsorption than that in other regions in Europe, Cd may pose higher environmental risks to this country. Lead, the metal with the highest soil adsorption capacity, might pose low environmental risks within a wide range of concentrations. However, Tóth *et al.* reported high percentages of soil samples with relatively high concentrations of Pb in Central Italy, France, Germany, and the UK, which could still of environmental concern.[86]

Different from Europe, countries in Africa are generally less developed and have less industrial processes involving heavy metals.[83] However, due to the lack of regulations as well as inadequate industrial waste monitoring and management capabilities, hazardous wastes are frequently released into the environment without treatment, which may cause heavy metal accumulation in soil and pose great environmental risks.[87] Yabe *et al.* reported that Cd and Pb are the most widespread heavy metals in Africa overall.[87] Considering the estimated relative soil adsorption capacities of the two metals in Africa (Figure 4), Cd might cause higher environmental risks. In northern Africa, the Egypt and Mediterranean coasts are reportedly polluted by municipal and industrial wastes associated with direct discharges.[87−89] Considering the higher soil adsorption capacities of the six metals than those in most other regions in Africa, the environmental risks in northern Africa might be lower than expected. Western Africa has heavy metal pollution in soils mainly due to petroleum extraction.[87] The corrosion of pipelines and discharges from oil industries in the Niger Delta region resulted in the pollution of crude oil, as well as Pb, Cd, Cu, Zn, and Cr in soil.[90] According to our prediction results, soils in western Africa have higher adsorption capacities than that in other regions in Africa, which indicated that the environmental risks may not be as high as expected. Southern Africa, the largest producer of gold in the world, has mining as the major source of heavy metal pollution.[87] Compared with other metals, Cu has an extremely high level in soil due to the extensive Cu mining in Zambia.[90] Given the estimated low soil relative adsorption capacities of Cu in this region, the corresponding environmental risks might be at higher levels. Eastern Africa has many waste dump sites, where a large amount of solid waste including industrial, agricultural, domestic, and medical wastes, are indiscriminately disposed.[87] The contents of metals such as Pb, Cd, Cr, Cu, and Zn in soil near the dump sites exceed the recommended limits, which may pose high environmental risks even though the soil adsorption capacities for these metals are high.

As the largest country in Asia, China has experienced rapid social and economic development in recent years, which inevitably results in heavy metal pollution at a large scale.[91] It is reported that more than 25% of total arable farmland in China has been contaminated by heavy metals such as Cd, Cr, Pb, and Zn.[83] In addition, regional differences are significant mainly due to variations in the industry and agriculture.[91] Overall, higher heavy metal concentrations are generally distributed in the southeast hills, the Yunnan-Guizhou Plateau, and the south part of the Yangtze River, where many mineral resources are located generating large amounts of heavy metals such as Cd, Cr, Pb, and Zn.[91] However, the environmental risks in these regions may not be as high as expected due to the high soil adsorption capacities (Figure 4). In contrast, the concentrations of heavy metals in northern China are generally low,[91,92] while the environmental risk may be higher than expected due to the low adsorption capacities of the soils.

Prediction for Equilibrium Concentrations of Heavy Metals. Many studies have examined heavy metal pollution in a single environmental medium (*e.g.*, soil, water, or air).[3] However, these media are constantly interacting with each

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

other; therefore, it is important to comprehensively evaluate heavy metal pollution in integrated systems.[93] For example, Tian *et al.* evaluated the heavy metal pollution in sediments and water in the coastal environments of both China and South Korea and quantified the interactions of heavy metals in the sediments and water.[93] However, the study was performed by collecting and analyzing both sediment and water samples, which is time-consuming and labor-intensive. Using the prediction models developed in this study, the heavy metal concentrations in sediments (*i.e.*, the adsorbed amounts) can be predicted once the concentrations in the surrounding water have been analyzed.

This process can also be reversed if a new model is developed to predict the concentrations in water once the soil/sediment samples have been analyzed. Toward this goal, using the same 4420 data points but switching the adsorption capacity and C_e data as the model input and output, respectively, we further developed 10 combined machine learning models following the same procedure as before. The R^2 and loss function values (Table S11) indicated that ET had the best model performance among the 10 combined models, with the R^2 of 0.735 and 0.780, rmse of 1.189 and 1.142, MAE of 0.878 and 0.875, Huber loss of 0.0829 and 0.0826, and Log-cosh loss of 0.4539 and 0.4417 for the validation data set and test data set, respectively. In addition, six independent models for the six heavy metals were developed with ET, which achieved better predictive performance than the combined model (Table S12). The good model performance suggested high accuracy of the reversed ET models for predicting C_e based on adsorbed amounts.

Environmental Implications. This study developed 10 machine learning models to predict the heavy metal adsorption on soils based on the properties of soils, solution systems, and heavy metals using 4420 experimental data points collected from 150 articles. After a comprehensive comparison based on R^2 and four loss functions, GBDT was found to be the best algorithm to produce accurate and chemically meaningful predictions. Based on the interpretation of the SHAP and MAS values, the importance of the involved features follows the order of CEC > C_e > pH (soil) > clay > OC > radius > hydra_radius > ratio > pH (solution) > the first IE > ionic strength > temperature. Six independent models with less inputs for each metal were further developed with the GBDT learning model and achieved better model performance than the combined model. With the independent models, the global distributions of soil relative adsorption capacities were predicted for all the six heavy metals based on the reported soil properties. The reversed models can also allow users to predict heavy metal concentrations in water when the adsorbed amount is known for a soil/sediment.

Compared with traditional risk assessment when only the heavy metal content is considered, the introduction of adsorption to heavy metal risk assessment is likely to provide more accurate results. The global distribution of the relative soil adsorption capacities obtained in this study is useful for facility location planning. The identified key soil properties influencing the adsorption capacities can help design soil remediation approaches to minimize the risks associated with heavy metals. The online predictor which is being developed in the adsorption section of ChemAI at https://www.chemai.aropha.com/ can make these models readily accessible for users with little programing skills. The sample python code can provide a step-by-step guidance for those who want to run the models with command-level control.

However, there are still some limitations in these models for real-world applications. The output of the main models only considers the adsorbed amounts based on the laboratory experiments, while in reality, the heavy metal contents in soils are likely influenced by other processes such as biological uptake and redox transformation. Therefore, the levels of heavy metals in soils/sediments may be underestimated using these models. For the same reason, the reversed ET model may overestimate the contents of heavy metals in water. To overcome such drawbacks, future studies should focus on collecting real-world data from water and soil/sediment integrated systems such that the adsorption and many other heavy metal transport/transformation processes are simultaneously considered.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.1c02479.

Additional details about the 10 machine learning algorithms and their performance; data-splitting approaches; *k*-fold cross validation; grid search; four regression loss functions; descriptor selection for heavy metal properties and their correlations; Shapely values and Shapley additive explanations; methods for similarity calculation; studies where the data points were extracted; effect of pH on heavy metal speciation; and mean relative soil adsorption capacities for different countries (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

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

Feier Wang − *College of Environmental & Resource Sciences, Zhejiang University, Hangzhou 310058, China;* Email: wangfeier@zju.edu.cn

Authors

Hongrui Yang − *College of Environmental & Resource Sciences, Zhejiang University, Hangzhou 310058, China*

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

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

Qin Weng − *College of Environmental & Resource Sciences, Zhejiang University, Hangzhou 310058, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c02479

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328

■ ACKNOWLEDGMENTS

H.Z. acknowledges the financial support by the National Science Foundation under grant # CHE-2105005. F.W. acknowledges the financial support by the National Major Science and Technology Program for Water Pollution Control and Treatment of China (2018ZX07208-009).

■ REFERENCES

(1) Vítková, M.; Rákosová, S.; Michálková, Z.; Komárek, M. Metal(loid)s behaviour in soils amended with nano zero-valent iron as a function of pH and time. *J. Environ. Manage.* **2017**, *186*, 268−276.
(2) Anagu, I.; Ingwersen, J.; Utermann, J.; Streck, T. Estimation of heavy metal sorption in German soils using artificial neural networks. *Geoderma* **2009**, *152*, 104−112.
(3) Yang, H.; Wang, F.; Yu, J.; Huang, K.; Zhang, H.; Fu, Z. An improved weighted index for the assessment of heavy metal pollution in soils in Zhejiang, China. *Environ. Res.* **2021**, *192*, 110246.
(4) Zhu, J.; Fu, Q.; Qiu, G.; Liu, Y.; Hu, H.; Huang, Q.; Violante, A. Influence of low molecular weight anionic ligands on the sorption of heavy metals by soil constituents: a review. *Environ. Chem. Lett.* **2019**, *17*, 1271−1280.
(5) Rosen, V.; Chen, Y. Effects of compost application on soil vulnerability to heavy metal pollution. *Environ. Sci. Pollut. Res.* **2018**, *25*, 35221−35231.
(6) Bolan, N.; Kunhikrishnan, A.; Thangarajan, R.; Kumpiene, J.; Park, J.; Makino, T.; Kirkham, M. B.; Scheckel, K. Remediation of heavy metal(loid)s contaminated soils−To mobilize or to immobilize? *J. Hazard. Mater.* **2014**, *266*, 141−166.
(7) Arunakumara, K. K. I. U.; Walpola, B. C.; Yoon, M.-H. Current status of heavy metal contamination in Asia's rice lands. *Rev. Environ. Sci. Biotechnol.* **2013**, *12*, 355−377.
(8) Imoto, Y.; Yasutaka, T. Comparison of the impacts of the experimental parameters and soil properties on the prediction of the soil sorption of Cd and Pb. *Geoderma* **2020**, *376*, 114538.
(9) Huang, B.; Yuan, Z.; Li, D.; Zheng, M.; Nie, X.; Liao, Y. Effects of soil particle size on the adsorption, distribution, and migration behaviors of heavy metal(loid)s in soil: a review. *Environ. Sci.: Processes Impacts* **2020**, *22*, 1596−1615.
(10) Elbana, T. A.; Selim, H. M.; Akrami, N.; Newman, A.; Shaheen, S. M.; Rinklebe, J. Freundlich sorption parameters for cadmium, copper, nickel, lead, and zinc for different soils: Influence of kinetics. *Geoderma* **2018**, *324*, 80−88.
(11) Azouzi, R.; Charef, A.; Hamzaoui, A. H. Assessment of effect of pH, temperature and organic matter on zinc mobility in a hydromorphic soil. *Environ. Earth Sci.* **2015**, *74*, 2967−2980.
(12) Alexakis, D. Diagnosis of stream sediment quality and assessment of toxic element contamination sources in East Attica, Greece. *Environ. Earth Sci.* **2011**, *63*, 1369−1383.
(13) Zhang, K.; Zhong, S.; Zhang, H. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, Granular activated carbons, and resins with machine learning. *Environ. Sci. Technol.* **2020**, *54*, 7008−7018.
(14) Buchter, B.; Davidoff, B.; Amacher, M. C.; Hinz, C.; Iskandar, I. K.; Selim, H. M. Correlation of Freundlich Kd and n retention parameters with soils and elements. *Soil Sci.* **1989**, *148*, 370−379.
(15) Bruemmer, G. W.; Gerth, J.; Herms, U. Heavy metal species, mobility and availability in soils. *Z. Pflanzenernähr. Bodenkd.* **1986**, *149*, 382−398.
(16) van der Zee, S. E. A. T. M.; Van Riemsdijk, W. H. Transport of reactive solute in spatially-variable soil systems. *Water Resour. Res.* **1987**, *23*, 2059−2069.
(17) Anderson, P. R.; Christensen, T. H. Distribution coefficients of Cd, Co, Ni, and Zn in soils. *J. Soil Sci.* **1988**, *39*, 15−22.
(18) Streck, T.; Richter, J. Heavy metal displacement in a sandy soil at the field scale .1. Measurements and parameterization of sorption. *J. Environ. Qual.* **1997**, *26*, 49−56.
(19) Schug, B.; Düring, R.-A.; Gäth, S. Improved cadmium sorption isotherms by the determination of initial contents using the radioisotope Cd-109. *J. Plant Nutr. Soil Sci.* **2000**, *163*, 197−202.
(20) Hudson, R. J. M.; Morel, F. M. M. Iron transport in merine-phytoplankton-kinetics of cellular and medium coordination reactions. *Limnol. Oceanogr.* **1990**, *35*, 1002−1020.
(21) Kraepiel, A. M. L.; Keller, K.; Morel, F. M. M. On the acid-base chemistry of permanently charged minerals. *Environ. Sci. Technol.* **1998**, *32*, 2829−2838.
(22) Kraepiel, A. M. L.; Keller, K.; Morel, F. M. M. A model for metal adsorption on montmorillonite. *J. Colloid Interface Sci.* **1999**, *210*, 43−54.
(23) Bradl, H. B. Adsorption of heavy metal ions on soils and soils constituents. *J. Colloid Interface Sci.* **2004**, *277*, 1−18.
(24) Gu, P.; Zhang, S.; Li, X.; Wang, X.; Wen, T.; Jehan, R.; Alsaedi, A.; Hayat, T.; Wang, X. Recent advances in layered double hydroxide-based nanomaterials for the removal of radionuclides from aqueous solution. *Environ. Pollut.* **2018**, *240*, 493−505.
(25) Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M.; Zou, X.; Wang, J.; Zhang, Y.; Chen, D.; Chen, X.; Deng, Y.; Ren, H. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* **2020**, *171*, 115454.
(26) 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.
(27) 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.
(28) 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.
(29) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, DOI: 10.1021/acs.est.1c01339.
(30) 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.
(31) 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*, 15385−15394.
(32) Felicísimo, Á. M.; Cuartero, A.; Remondo, J.; Quirós, E. Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslides* **2013**, *10*, 175−189.
(33) Patil, M. A.; Tagade, P.; Hariharan, K. S.; Kolake, S. M.; Song, T.; Yeo, T.; Doo, S. A novel multistage Support Vector Machine based approach for Li ion battery remaining useful life estimation. *Appl. Energy* **2015**, *159*, 285−297.
(34) Tian, J.; Morillo, C.; Azarian, M. H.; Pecht, M. Motor Bearing Fault Detection Using Spectral Kurtosis-Based Feature Extraction Coupled With K-Nearest Neighbor Distance Analysis. *IRE Trans. Ind. Electron.* **2016**, *63*, 1793−1803.
(35) Qiu, X.; Ren, Y.; Suganthan, P. N.; Amaratunga, G. A. J. Empirical mode decomposition based ensemble deep learning for load demand time series forecasting. *Appl. Soft Comput.* **2017**, *54*, 246−255.

(36) Tyralis, H.; Papacharalampous, G.; Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* **2019**, *11*, 910.
(37) Sahin, E. K. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Appl. Sci.* **2020**, *2*, 1308.
(38) Bhagat, S. K.; Tiyasha, T.; Awadh, S. M.; Tung, T. M.; Jawad, A. H.; Yaseen, Z. M. Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models. *Environ. Pollut.* **2021**, *268*, 115663.
(39) Bazoobandi, A.; Emamgholizadeh, S.; Ghorbani, H. Estimating the amount of cadmium and lead in the polluted soil using artificial intelligence models. *Eur. J. Environ. Civ. Eng.* **2019**, 1.
(40) Tan, K.; Wang, H.; Chen, L.; Du, Q.; Du, P.; Pan, C. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. *J. Hazard. Mater.* **2020**, *382*, 120987.
(41) Jia, X.; Hu, B.; Marchant, B. P.; Zhou, L.; Shi, Z.; Zhu, Y. A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: A case study in the Yangtze Delta, China. *Environ. Pollut.* **2019**, *250*, 601−609.
(42) Kokkinos, Y.; Margaritis, K. G. Managing the computational cost of model selection and cross-validation in extreme learning machines via Cholesky, SVD, QR and eigen decompositions. *Neurocomputing* **2018**, *295*, 29−45.
(43) Nightingale, E. R. Phenomenological theory of ion solvation. effective radii of hydrated ions. *J. Phys. Chem.* **1959**, *63*, 1381−1387.
(44) Qiang, S.; Han, B.; Zhao, X.; Yang, Y.; Shao, D.; Li, P.; Liang, J.; Fan, Q. Sorption of nickel(II) on a calcareous aridisol soil, China: Batch, XPS, and EXAFS spectroscopic investigations. *Sci. Rep.* **2017**, *7*, 46744.
(45) Chaib, A.; Boukhalfa, N.; Boudjemaa, A. Congo Red removal using Fesdis soil: kinetic, equilibrium and thermodynamic studies. *Int. J. Environ. Anal. Chem.* **2021**, 1.
(46) Jalali, M.; Vafaee, Z.; Fakhri, R. Selectivity Sequences of Heavy Metals in Single and Competitive Systems under Different Soil/Solution Ratios and pH in a Calcareous Soil. *Commun. Soil Sci. Plant Anal.* **2020**, *51*, 341−351.
(47) Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 1−21.
(48) Shapley, L. S. A value for n-person game. In *Contributions to the Theory of Games*; Kuhn, H. W., Tucker, A. W., Eds.; Princeton University Press: Princeton, 1953; pp 307−318.
(49) Zhong, L.; Guo, X.; Xu, Z.; Ding, M. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. *Geoderma* **2021**, *402*, 115366.
(50) Fischer, G.; Nachtergaele, F.; Prieler, S.; Velthuizen, H. T. V.; Verelst, L.; Wiberg, D. Harmonized World Soil Database v1.2 (GAEZ, IIASA & FAO). http://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/, 2008.
(51) Ma, L.; Xu, R.; Jiang, J. Adsorption and desorption of Cu(II) and Pb(II) in paddy soils cultivated for various years in the subtropical China. *J. Environ. Sci.* **2010**, *22*, 689−695.
(52) McBride, M.; Sauve, S.; Hendershot, W. Solubility control of Cu, Zn, Cd and Pb in contaminated soils. *Eur. J. Soil Sci.* **1997**, *48*, 337−346.
(53) Hooda, P. S.; Alloway, B. J. Cadmium and lead sorption behaviour of selected English and Indian soils. *Geoderma* **1998**, *84*, 121−134.
(54) Nakamura, K.; Yasutaka, T.; Kuwatani, T.; Komai, T. Development of a predictive model for lead, cadmium and fluorine soil−water partition coefficients using sparse multiple linear regression analysis. *Chemosphere* **2017**, *186*, 501−509.
(55) Sauvé, S.; McBride, M. B.; Norvell, W. A.; Hendershot, W. H. Copper solubility and speciation of in situ contaminated soils: Effects of copper level, pH and organic matter. *Water, Air, Soil Pollut.* **1997**, *100*, 133−149.
(56) Xu, L.; Wang, T.; Luo, W.; Ni, K.; Liu, S.; Wang, L.; Li, Q.; Lu, Y. Factors influencing the contents of metals and As in soils around the watershed of Guanting Reservoir, China. *J. Environ. Sci.* **2013**, *25*, 561−568.
(57) Ugwu, I. M.; Igbokwe, O. A. Sorption of heavy metals on clay minerals and oxides: a review. *Advanced Sorption Process Applications*; IntechOpen, 2019.
(58) Baker, J. F.; Burrows, N. L.; Keohane, A. E.; de Filippis, L. F. Chemical root pruning of kangroo paw (anigozanthos-flaidus) by selected heavy-metal carbonates. *Sci. Hortic.* **1995**, *62*, 245−253.
(59) Pignatello, J. J.; Xing, B. Mechanisms of slow sorption of organic chemicals to natural particles. *Environ. Sci. Technol.* **1996**, *30*, 1−11.
(60) Brallier, S.; Harrison, R. B.; Henry, C. L.; Xue, D. Liming effects on availability of Cd, Cu, Ni and Zn in a soil amended with sewage sludge 16 years previously. *Water, Air, Soil Pollut.* **1996**, *86*, 195−206.
(61) Rieuwerts, J. S.; Thornton, I.; Farago, M. E.; Ashmore, M. R. Factors influencing metal bioavailability in soils: preliminary investigations for the development of a critical loads approach for metals. *Chem. Speciat. Bioavailab.* **1998**, *10*, 61−75.
(62) Basta, N. T.; Tabatabai, M. A. Effect of cropping systems on adsorption of metals by soils. 2. effect of pH. *Soil Sci.* **1992**, *153*, 195−204.
(63) Zhang, C.; Gu, X.; Gu, C.; Evans, L. J. Multi-surface modeling of Ni(II) and Cd(II) partitioning in soils: Effects of salts and solid/liquid ratios. *Sci. Total Environ.* **2018**, *635*, 859−866.
(64) Mizutani, K.; Fisher-Power, L. M.; Shi, Z.; Cheng, T. Cu and Zn adsorption to a terrestrial sediment: Influence of solid-to-solution ratio. *Chemosphere* **2017**, *175*, 341−349.
(65) Bordas, F.; Bourg, A. Effect of solid/liquid ratio on the remobilization of Cu, Pb, Cd and Zn from polluted river sediment. *Water, Air, Soil Pollut.* **2001**, *128*, 391−400.
(66) Fotovat, A.; Naidu, R.; Sumner, M. E. Water:soil ratio influences aqueous phase chemistry of indigenous copper and zinc in soils. *Aust. Soil Res.* **1997**, *35*, 687−709.
(67) Naidu, R.; Bolan, N. S.; Kookana, R. S.; Tiller, K. G. Ionic-strength and pH effects on the sorption of cadmium and the surface-charge on soils. *Eur. J. Soil Sci.* **1994**, *45*, 419−429.
(68) Orumwense, F. F. O. Removal of lead from water by adsorption on a kaolinitic clay. *J. Chem. Technol. Biotechnol.* **1996**, *65*, 363−369.
(69) Liu, A.; Gonzalez, R. D. Adsorption/desorption in a system consisting of humic acid, heavy metals, and clay minerals. *J. Colloid Interface Sci.* **1999**, *218*, 225−232.
(70) Xu, R.; Wang, Y.; Tiwari, D.; Wang, H. Effect of ionic strength on adsorption of As(III) and As(V) on variable charge soils. *J. Environ. Sci.* **2009**, *21*, 927−932.
(71) Davari, M.; Rahnemaie, R.; Homaee, M. Competitive adsorption-desorption reactions of two hazardous heavy metals in contaminated soils. *Environ. Sci. Pollut. Res.* **2015**, *22*, 13024−13032.
(72) Dong, D.; Li, Y.; Zhang, J.; Hua, X. Comparison of the adsorption of lead, cadmium, copper, zinc and barium to freshwater surface coatings. *Chemosphere* **2003**, *51*, 369−373.
(73) Chu, Z.; Gu, W.; Li, Y. Adsorption mechanism of heavy metals in heavy metal/pesticide coexisting sediment systems through factional factorial design assisted by 2D-QSAR models. *Pol. J. Environ. Stud.* **2018**, *27*, 2451−2461.
(74) Faur-Brasquet, C.; Kadirvelu, K.; Le Cloirec, P. Removal of metal ions from aqueous solution by adsorption onto activated carbon cloths: adsorption competition with organic matter. *Carbon* **2002**, *40*, 2387−2392.
(75) Kong, Q.; Preis, S.; Li, L.; Luo, P.; Wei, C.; Li, Z.; Hu, Y.; Wei, C. Relations between metal ion characteristics and adsorption performance of graphene oxide: A comprehensive experimental and theoretical study. *Sep. Purif. Technol.* **2020**, *232*, 115956.
(76) Akpomie, K. G.; Dawodu, F. A.; Adebowale, K. O. Mechanism on the sorption of heavy metals from binary-solution by a low cost montmorillonite and its desorption potential. *Alexandria Eng. J.* **2015**, *54*, 757−767.

(77) Sierra, A.; Cruz, C. S. Global and local neural network ensembles. *Pattern Recogn. Lett.* **1998**, *19*, 651−655.
(78) Kawaguchi, K.; Huang, J.; Kaelbling, L. P. Effect of Depth and Width on Local Minima in Deep Learning. *Neural Comput.* **2019**, *31*, 1462−1498.
(79) Li, Y.; Ngom, A. Nonnegative Least-Squares Methods for the Classification of High-Dimensional Biological Data. *IEEE ACM Trans. Comput. Biol. Bioinf.* **2013**, *10*, 447−456.
(80) Zhu, X.; Zhang, S.; Hu, R.; Zhu, Y.; Song, J. Local and Global Structure Preservation for Robust Unsupervised Spectral Feature Selection. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 517−529.
(81) Frachini, E.; Constantino, L. V.; Abrao, T.; Santos, M. J. A new approach to evaluate toxic metal transport in a catchment. *Environ. Monit. Assess.* **2020**, *192*, 234.
(82) Hou, D.; O'Connor, D.; Igalavithana, A. D.; Alessi, D. S.; Luo, J.; Tsang, D. C. W.; Sparks, D. L.; Yamauchi, Y.; Rinklebe, J.; Ok, Y. S. Metal contamination and bioremediation of agricultural soils for food safety and sustainability. *Nat. Rev. Earth Environ.* **2020**, *1*, 366−381.
(83) Khalid, S.; Shahid, M.; Niazi, N. K.; Murtaza, B.; Bibi, I.; Dumat, C. A comparison of technologies for remediation of heavy metal contaminated soils. *J. Geochem. Explor.* **2017**, *182*, 247−268.
(84) Foucault, Y.; Lévêque, T.; Xiong, T.; Schreck, E.; Austruy, A.; Shahid, M.; Dumat, C. Green manure plants for remediation of soils polluted by metals and metalloids: Ecotoxicity and human bioavailability assessment. *Chemosphere* **2013**, *93*, 1430−1435.
(85) Goix, S.; Lévêque, T.; Xiong, T.-T.; Schreck, E.; Baeza-Squiban, A.; Geret, F.; Uzu, G.; Austruy, A.; Dumat, C. Environmental and health impacts of fine and ultrafine metallic particles: Assessment of threat scores. *Environ. Res.* **2014**, *133*, 185−194.
(86) Tóth, G.; Hermann, T.; Da Silva, M. R.; Montanarella, L. Heavy metals in agricultural soils of the European Union with implications for food safety. *Environ. Int.* **2016**, *88*, 299−309.
(87) Yabe, J.; Ishizuka, M.; Umemura, T. Current levels of heavy metal pollution in Africa. *J. Vet. Med. Sci.* **2010**, *72*, 1257−1263.
(88) Abdallah, M. A. M. Trace metal behaviour in Mediterranean-climate coastal bay: El-Mex Bay, Egypt and its coastal environment. *Global J. Environ. Res.* **2008**, *2*, 23−29.
(89) El-Rayis, O. A.; Abdallah, M. A. M. Contribution of nutrients and some trace metals from a huge Egyptian Drain to the SE-Mediterranean Sea, west of Alexandria. *Mediterr. Mar. Sci.* **2006**, *7*, 79−86.
(90) Chindah, A. C.; Braide, S.; Sibeudu, O. C. Distribution of hydrocarbons and heavy metals in sediment and a crustacean (shrimps - Penaeus notialis) from the Bonny/New Calabar River Estuary, Niger Delta. *African Journal of Environmental Assessment and Management*, 2004; Vol. *9*, pp 1−17.
(91) Yang, Q.; Li, Z.; Lu, X.; Duan, Q.; Huang, L.; Bi, J. A review of soil heavy metal pollution from industrial and agricultural regions in China: Pollution and risk assessment. *Sci. Total Environ.* **2018**, *642*, 690−700.
(92) Zhang, X.; Yang, L.; Li, Y.; Li, H.; Wang, W.; Ye, B. Impacts of lead/zinc mining and smelting on the environment and human health in China. *Environ. Monit. Assess.* **2012**, *184*, 2261−2273.
(93) Tian, K.; Wu, Q.; Liu, P.; Hu, W.; Huang, B.; Shi, B.; Zhou, Y.; Kwon, B.-O.; Choi, K.; Ryu, J.; Seong Khim, J.; Wang, T. Ecological risk assessment of heavy metals in sediments and water from the coastal areas of the Bohai Sea and the Yellow Sea. *Environ. Int.* **2020**, *136*, 105512.

https://doi.org/10.1021/acs.est.1c02479
Environ. Sci. Technol. 2021, 55, 14316−14328