Direct prediction of bioaccumulation of organic contaminants in plant roots from soils with machine learning models based on molecular structures

Feng Gao
Yale University
Yike Shen
Columbia University
J. Brett Sallach
University of York
Hui Li
Michigan State University
Cun Liu (liucun@issas.ac.cn)
Institute of Soil Science, Chinese Academy of Sciences
Yuanbo Li
Institute of Plant Protection, Chinese Academy of Agricultural Sciences

Article

Keywords: root concentration factor, organic contaminants, soils

Posted Date: March 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-240794/v1

Version of Record: A version of this preprint was published at Environmental Science & Technology on December 3rd, 2021. See the published version at https://doi.org/10.1021/acs.est.1c02376.

Direct prediction of bioaccumulation of organic contaminants in plant roots from soils with machine learning models based on molecular structures

Feng Gao[1], Yike Shen[2], J. Brett Sallach[3], Hui Li[4], Cun Liu[5]*, Yuanbo Li[6]*

[1] Department of Genetics, School of Medicine, Yale University, New Haven, Connecticut 06510, United States

[2] Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, New York 10032, United States

[3] Department of Environment and Geography, University of York, Heslington, York, YO10 5NG, United Kingdom

[4] Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, Michigan, 48823, United States

[5] Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, P.R. China

[6] State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100193, P.R. China

* Corresponding authors. Tel.: 86-25-86881179, fax: 86-25-86881000 (C Liu); Tel.: 86-10-62815938, fax: 86-10-62896114 (Y Li)

E-mail address: liucun@issas.ac.cn (C Liu), liyuanbo@caas.cn (Y Li).

Abstract

Root concentration factor (RCF_{soil}) is an important substance-specific characterization parameter for plant uptake of organic contaminants from soils in life cycle impact assessment (LCIA); however, the availability of a reliable dataset and building of robust predictive models remain challenging due to the complexity of chemical-soil-plant root interactions. Here we developed end-to-end machine learning models to devolve the interaction complexity by training on a unified RCF_{soil} dataset with 341 data points covering 72 chemicals. The gradient boosting regression tree (GBRT) model based on the extended connectivity fingerprints (ECFP) demonstrated a superior RCF_{soil} prediction performance with R-squared of 0.77 and Mean Absolute Error (MAE) of 0.22. In addition, partial dependence analysis was used to determine the nonlinear relationships in the chemical-soil-plant root system. Feature importance analysis revealed the relationship between RCF_{soil} and chemical topological structures. Stemming from its simplicity and universality, the GBRT-ECFP model provides a promising tool for LCIA to better characterize the human and ecological impacts of chemicals in the environment.

Main

The rapid increase in production and usage of diverse groups of anthropogenic organic chemicals inevitably results in their release into the environment where soil is the major sink for many categories of organic contaminants[1,2]. The transport of organic contaminants from soil into plants, especially crops, poses a great threat to agricultural sustainability and potentially harms human health through dietary consumption[3]. Laboratory and field studies have been conducted to measure the transfer of various organic chemicals from soils to specific plants[4]. However, the rapidly increasing amount of existing and new chemicals are not able to be evaluated experimentally in a timely manner. This has necessitated the development of reliable prediction

tools to evaluate the transfer of contaminants from soil to plants based on molecular signature of chemicals[5,6]. These tools could be utilized in life cycle impact assessments (LCIA) for organic chemicals within their life cycle assessment (LCA) framework[7].

As an essential substance-specific characterization parameter in LCIA, the accumulation of environmental contaminants in plant roots is routinely evaluated in many experimental and modeling studies by root concentration factor (RCF). RCF is defined as the ratio of contaminant concentration in roots to that in soil at a steady state or equilibrium state among compartments including soil solids, soil solution (interstitial water) and the plant roots[5,8,9]. RCF is controlled by chemical, plant and soil properties, as a result of sorption–desorption processes at the interface of water-soil components, including soil organic matter (OM) and clay minerals, and partitioning processes at the water-plant root system that is influenced by lipid and carbohydrate contents. Compared to RCF based on water concentration RCF_{water} that can be well predicted by empirical models, RCF directly based on soil concentration RCF_{soil} is a more relevant and measurable characterization factor in LCIA. However, it is difficult to accurately estimate the RCF_{soil} value due to the complexity of multiple interactions for contaminants at the interfaces of soils, water, plant roots and microorganisms (i.e., rhizosphere). Models for the direct accurate prediction of RCF_{soil} are lacking, as most available empirical models have reported relative uncertainties for RCF_{soil} predictions around one order of magnitude[10]. Furthermore, the relationships between the chemical structures of contaminants and their uptake mechanisms from soil is not well understood.

Machine learning and deep learning algorithms have been widely used in image recognition, natural language processing, and with chemistry applications including reaction prediction and molecular property prediction[11,12]. Recently, as big data-based assessment and

decision-making tools, machine learning models were successfully applied to predict some characterization parameters of LCIA, such as chemical USEtox HC50 values [13-15]. However, applying these algorithms to other LCIA characterization parameters, such as RCF_{soil} in the plant-soil system, remain challenging. This not only requires computational models to be able to learn complex relationships among multiple variables but also depends on the selection of well representative variables involved. Similar to the traditional empirical models, e.g., compartment models and quantitative structure-activity relationship (QSAR) models, common representations could be physicochemical properties [16-18]. For example, adsorption predictions of organic molecules on carbonaceous materials and polymers have used molecular physicochemical properties such as Abraham descriptors to represent the molecules[16,18]. However, these physicochemical properties are usually selected based on the assumptions of their importance in sorption processes and do not adequately account for molecular structure. Therefore, a more objective approach will be encoded with fundamental molecular structure information, rather than the general physicochemical properties.

Molecular fingerprint is a way of encoding molecular structures with a series of binary bits to describe the presence or absence of a certain chemical structure. One of the most commonly used fingerprint is Extended Connectivity Fingerprints (ECFP)[19] which is a type of circular topological fingerprints. ECFP is developed specifically for structure-activity modeling, not predefined, and can represent an essentially infinite number of different molecular features (including stereochemical information). These molecular features represent the presence of particular substructures in the molecule, allowing easier interpretation of analysis results. By replacing molecular physicochemical properties with ECFP, it becomes feasible to directly relate RCF_{soil} with chemical structures, which provides more rich information about how structures

could affect RCF_{soil}, while avoiding ambiguous selection of molecular physicochemical properties based on assumptions for the model.

In this study, the uptake of organic chemicals by plants from soils was predicted by machine learning methods based on the collection of reliable RCF_{soil} data from literatures. ECFP were used in conjunction with gradient boosting decision tree (GBRT) methods for the development of unbiased predictive models. The workflow from data collection (Fig. 1a), construction of machine learning models (Fig. 1b-d) to systematic evaluation (Fig. 1e) and application in LCIA (Fig. 1f) is constructed. The contributions of this study are three-fold: 1) the established GBRT-ECFP model achieved the state-of-the-art prediction results for uptake of organic chemicals by plants from soil in terms of R-squared value and significantly reduced relative uncertainty for RCF_{soil} prediction; 2) GBRT-ECFP model did not rely on the molecular physicochemical properties acquired from experiments thus has potentially wider applications; and 3) the feature importance analysis revealed the relationships between molecular substructures and RCF_{soil}. This study facilitates a deeper understanding of plant uptake mechanisms and allows these findings to be extended to novel molecules in development or pre-registration phase. This new machine learning model can be used to explore structure-related information for screening large number of emerging chemicals and provide a simple and reliable tool for LCIA to better characterize the ecotoxicological impacts of chemicals in the environment.

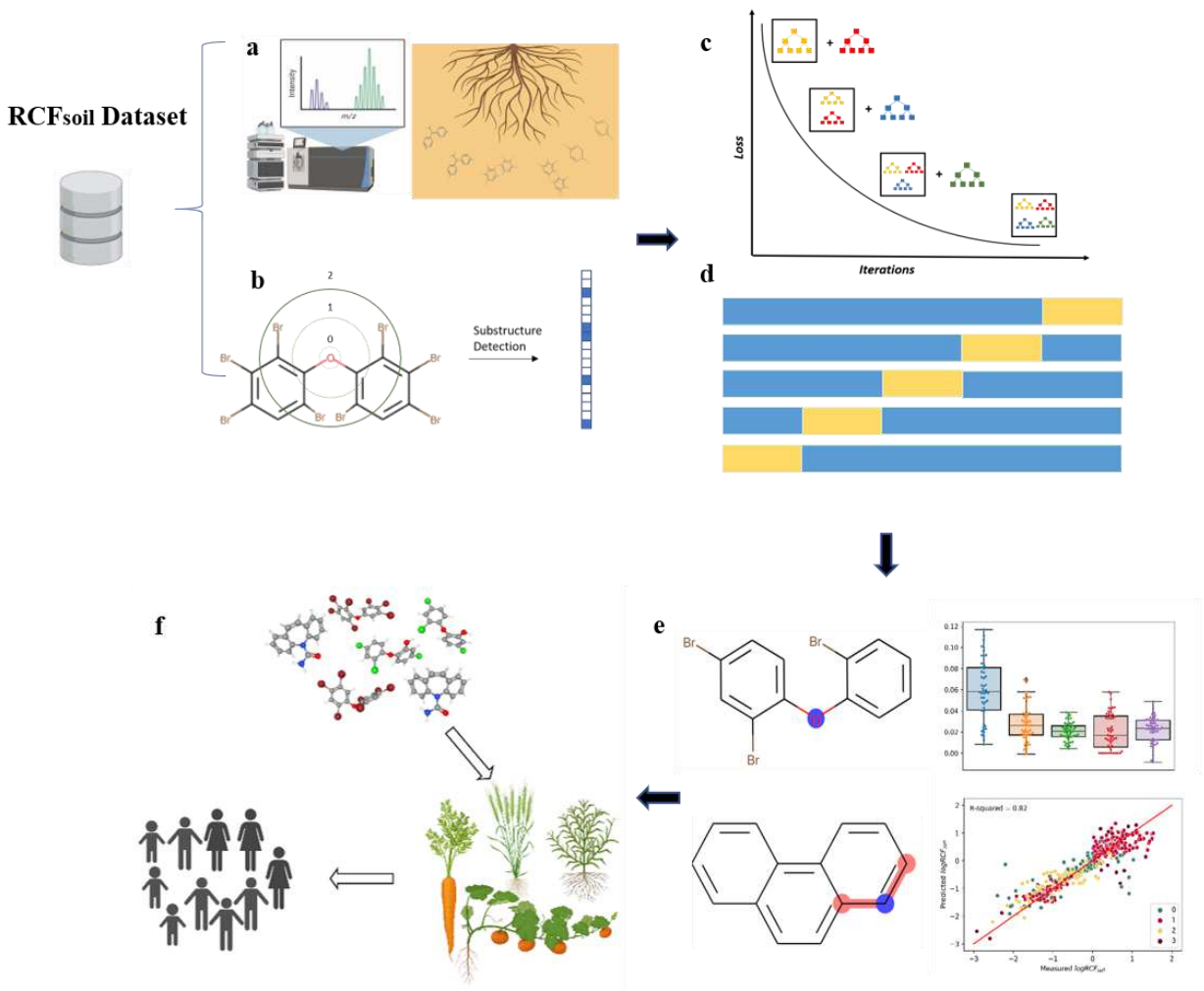

Fig. 1: Schematic representation for building machine learning model for RCF_{soil} prediction and implications: **a**, Data collected from plant uptake studies**. b**, ECFP used for molecular structures representation. **c**, Gradient boosting regression tree model trained on the dataset. **d**, 5-fold cross validation for parameter tuning. **e**, RCF_{soil} prediction and model interpretation. **f**, Implication of model results in LCIA.

Results and discussion

Molecular representation with ECFP

RCF_{soil} is believed to closely related to chemical structures, hence, we started from analyzing the structure variation of chemicals in the dataset. The structures of molecule pairs were compared based on their ECFP fingerprints. An example of this molecular pair comparison of 2,2',4,4'-tetrabromodiphenyl ether (BDE-47) with fluoranthene, 2,2',3,3',4,4',6,6'-octabromodiphenyl ether (BDE-197), decabromodiphenyl ether is (BDE-209) shown in Fig. 2a. Similar substructures were marked in green and the differences in purple. These pair comparison patterns showed that ECFP can effectively distinguish different substructures among molecules.

To further demonstrate the validity of ECFP fingerprint for molecular representation, the 72 chemicals collected in the dataset were clustered into groups based on their ECFP representation using the K-means clustering algorithm (Fig. 2b). Molecules with similar structures were clustered into four groups, and the list of detailed chemicals in each group can be found in SI Table S1. In summary, Group 0 contains 32 chemicals made up of chlorinated chemicals such as PCBs, polychlorobenzenes, pesticides, and several other chemicals including pharmaceuticals and personal care products (PPCPs) and fragrances. Group 1 contains 16 chemicals, which is solely composed of polybromodiphenyl ethers (PBDEs). Group 2 contains 15 chemicals almost exclusively PAHs, and Group 3 contains 9 chemicals, which are pesticides, PPCPs and other chemicals slightly overlap with other groups. The clear classification of diverse groups of chemicals based on the unbiased ECFP descriptions for molecular structures provided further evidence of the effectiveness of using ECFP as molecular structure descriptors.

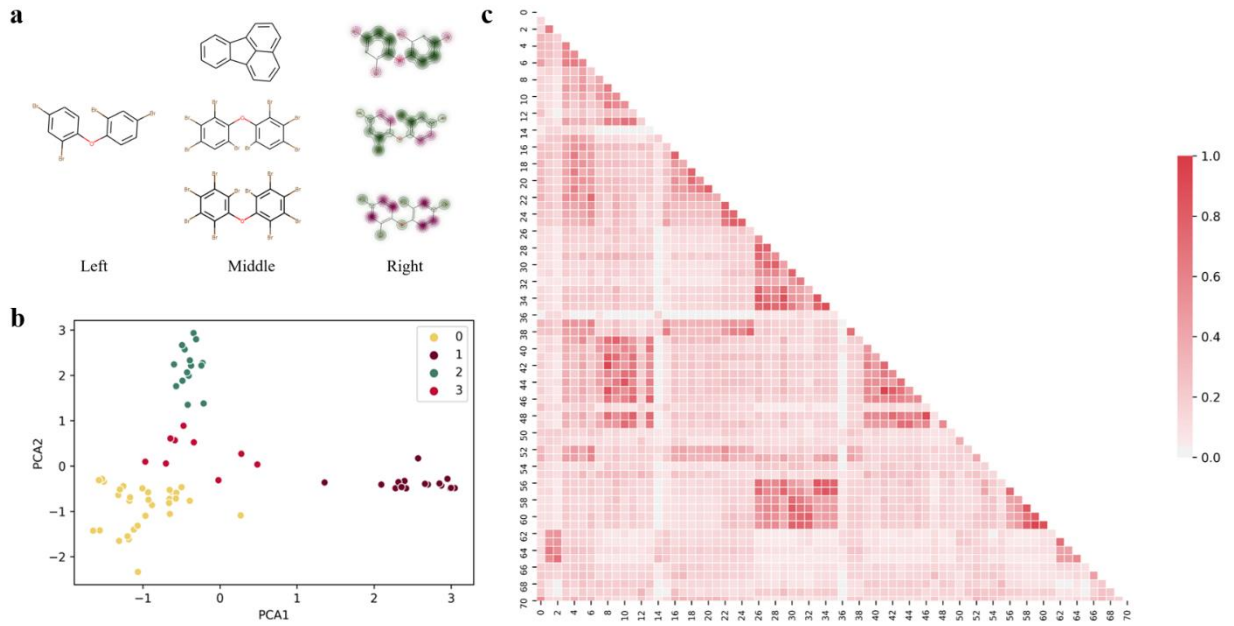

Fig. 2: **a**, Comparison of 2,2’,4,4’-tetrabromodiphenyl ether (BDE-47, left) with three molecules (middle) from top to bottom: fluoranthene, 2,2’,3,3’,4,4’,6,6’-octabromodiphenyl ether (BDE-197), decabromodiphenyl ether (BDE-209). Comparison results (right) with dark green showing similar substructures while dark red showing different substructures. **b**, Clustering of chemicals based on ECFP fingerprints. **c**, Similarity comparison of chemicals in the dataset (darker red shows higher similarities).

An overall comparison of molecule structure similarity using ECFP is shown in Fig. 2c. Dice similarity, a feature-based similarity coefficient, was calculated for each pair of molecules. The structure similarity reflected by dice similarity scores ranged from 0.0 to 1.0 (More details can be found in SI Table S2). Chemicals with similar molecular structures are supposed to have a higher score than structurally dissimilar chemicals. For example, 2,2',3,4,4',5',6-heptabromodiphenyl ether (BDE-183) and 2,2',3,3',4,5,5',6,6'-nonabromodiphenyl ether (BDE-208) share similar structures and thus has a score of 0.67, while Aldrin and 2,2',3,3',4,5,5',6,6'-nonabromodiphenyl ether (BDE-208) has a dice similarity score of 0.11 (chemical structures of the three chemicals can be found in SI Fig. S1). In addition, the comparison showed that the

dataset covered a large variation of different chemical structures, which makes the dataset potentially representative for evaluating different machine learning models on RCF_{soil} prediction.

$logRCF_{soil}$ prediction with GBRT

For comparison, the linear-regression (LR) model with representative descriptors: octanol-water partition coefficient ($logKow$), molecular weight (MW), soil organic matter content (f_{OM}) and plants' root lipid content (f_{lipid}) was used as benchmark to the GBRT model. It should be noted that the linear regression models implicitly assume a linear relationship between $logRCF$ and descriptor variables. The LR results, predicted values versus measured values, are shown in Fig. 3a. The same property descriptors were also used as inputs to the GBRT model (results shown in Fig. 3 b). All the features were z-normalized before feeding to the models, and different colors represent groupings from the previous clustering results. Our results showed that the GBRT model achieved improved prediction results with R-squared of 0.73 and mean absolute error (MAE) of 0.24 compared to the inferior LR model with R-squared of 0.60 and the MAE of 0.35.

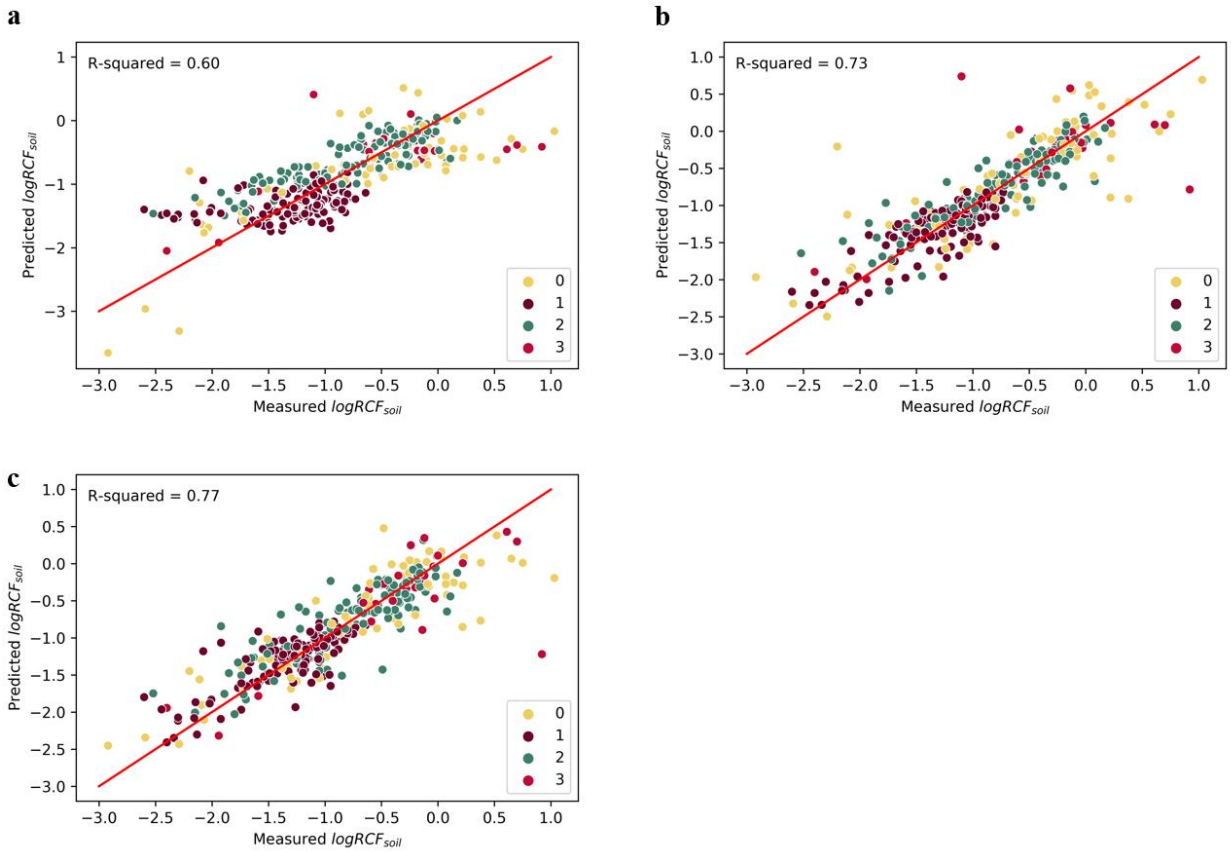

Fig. 3: Prediction of $logRCF_{soil}$ with **a**, LR model and molecular physicochemical property descriptors; **b**, GBRT model and molecular physicochemical property descriptors; and **c**, ECFP-GBRT model. Different colors represent different chemical groups from clustering.

The poor performance of the LR model is indicative of the complex relationship between $logRCF_{soil}$ and physicochemical properties of the contaminants as well as plant and environmental parameters such as the fraction of organic matter f_{OM}, and lipid content of plant roots f_{lipid}. Similarly, in a recent study, Li et.al reported that no apparent correlation was observed between $logK_{ow}$ and $logRCF_{soil}$ for 374 data points summarized across 19 soil-plant studies [20]. More complicated models for plant uptake of organic contaminants have been developed and attempted to incorporate the complexity of uptake processes into their formulation by integrating specific uptake pathways, compound classes, and/or plant types [2]. Although such

compartment models explicitly include a more complete set of uptake processes, they still suffer from poor predictive accuracy. Collins *et al.* examined nine models for root uptake of nonionizable contaminants by comparing predictions to experimental study results including a wide range of chemical properties and uptake pathways and observed that most models over-predicted the contaminant concentrations in roots by at least one order of magnitude[10]. This further confirmed that for complex systems (e.g., soil-plant system), current knowledge on the uptake mechanisms were not adequate for reliable prediction of plant root uptake. Moreover, LR model is limited by its inability to capture nonlinear relationship. GBRT model, with same property descriptor inputs significantly improved the prediction accuracy. This is an indication that the complex relationship between $logRCF_{soil}$ and chemicals, soil and root properties cannot be simply assumed as linear uptake and bioconcentration. Indeed, further analysis was conducted on how the changes of property descriptors affected the predicted $logRCF_{soil}$ through partial dependence plot. The influences of $logK_{ow}$, f_{OM}, MW, $and\ f_{lipid}$ in terms of their z-scored values are shown in SI Fig S2. The partial dependence plots revealed nonlinear relationships between multiple known and unknown competing pathways within plant compartments and among the soil-plant continuum. This explains why LR and compartment models are not able to capture the complex relationships.

$logRCF_{soil}$ prediction with GBRT-ECFP model

Molecular physicochemical properties ($logKow$ and MW) were replaced by ECFP for $logRCF_{soil}$ prediction with the GBRT model. ECFP fingerprints were then concatenated with plant f_{lipid} and soil f_{OM} to form new features for each data point. The results for the GBRT-ECFP model are shown in Fig. 3c. By utilizing ECFP as molecular representation, the R-squared was further increased to 0.77 and MAE decreased to 0.22. ECFP captures layered atom

environments in compounds up to a pre-determined bond diameter (radius of 2 in this study), and the enhanced model performance when using ECFP compared to using property descriptors shows that molecular structures better predict $logRCF_{soil}$. The usage of ECFP can be helpful to capture potential interactions beyond hydrophobicity reflected by $logKow$. For example, when the MAE for all compounds containing nitrogen atoms was considered (a total of 15 data points), the MAE is 0.37 for those using ECFP and 0.44 for those using property descriptors. For all compounds that contain oxygen atoms (a total of 157 data points), the MAE for using ECFP is 0.21 slightly better than using physicochemical properties which is 0.24.

Another advantage of using ECFP fingerprints is its ability to distinguish isomers that have similar $logKow$ and MW but significantly different plant uptake behavior due to different structures[21]. The prediction of regioselective properties and reactions are quite challenging because they are also dependent upon the spatial position of the functional groups [22,23]. For example, pesticides o,p'-DDE and p,p'-DDE are isomers with different spatial position of chlorine substitutions and their $logKow$ are 5.76 and 5.91, respectively. However, their RCF_{soil} varied by almost 50%. The RCF_{soil} is 0.032 ($logRCF_{soil}$ -1.49) for o,p' DDE and 0.058 (logRCF_{soil} -1.24) for p'p-DDE under the same experimental conditions. This highlights the importance of stereochemistry in biological processes that largely neglected by scalar descriptors. With topological fingerprints, the prediction results of o,p' DDE ($logRCF_{soil}$ -1.49) were similar (ECFP fingerprint: -1.29 vs property descriptors: -1.31). However, the prediction results for p',p DDE (logRCF_{soil} -1.24) was improved from -1.59 (property descriptors) to -1.21 (ECFP), providing more accurate predictions.

Model interpretation by substructure analysis

Interpretability of machine learning models is one of the key challenges for their wide applications in environmental scenarios. ECFP accounts for the presence of particular substructures in the chemicals, and these substructures contribute differentially to the model. With the GBRT-ECFP model, it is now possible to relate substructures of chemicals with the output result, in this case RCF_{soil}. As far as we know, this is the first study to analyze how molecular substructures affect the RCF_{soil}. To better interpret the proposed GBRT-ECFP model and identify important substructures related to RCF_{soil}, two different feature importance analysis methods: permutation feature importance analysis and impurity feature importance analysis were applied. Detailed description of these importance analyses is provided in the SI. Briefly, permutation feature importance systematically identifies key features (substructures) through individual permutation of each input feature to evaluate its influence on the $logRCF_{soil}$ prediction results. The larger the effects, the higher the importance. To avoid the randomness in permuting features, each input feature was shuffled 10 times and the average were used to evaluate the feature importance. Impurity importance is another feature importance analysis method specific to tree-based methodologies.

The top five most important substructures found in this dataset identified by permutation importance are shown in Fig. 4a (chemicals composed of important substructure selected randomly as exemplars). The results showed that functional groups such as -O and -Cl and benzene rings or large π systems are important substructures that contribute to the prediction of $logRCFs$, which is closely related to hydrophobicity or lipophilicity of chemicals. It matches previous studies that $logKow$ contributes predominantly to the prediction of $logRCFs$[12], while f_{lipid} and f_{OM} consider the competitive sorption between soil organic carbon and plant lipids[24]. Similarly in Fig. 4b, the top five most important substructures identified by impurity-based

importance method were also related to -O, -Cl, benzene rings and large π systems. Other topological features, though less important, substantially improved the prediction accuracy by comprehensively accounting for other potential interactions. The consistency of identifying similar structures with two different techniques (permutation importance analysis and impurity importance analysis) is further proof of the reliability of the proposed GBRT-ECFP model. Most importantly, the machine learning method utilizing unbiased topological molecular fingerprints was able to reveal the most relevant molecular structural features of organic contaminants for a more accurate prediction of the chemical uptake by plant roots from soil.

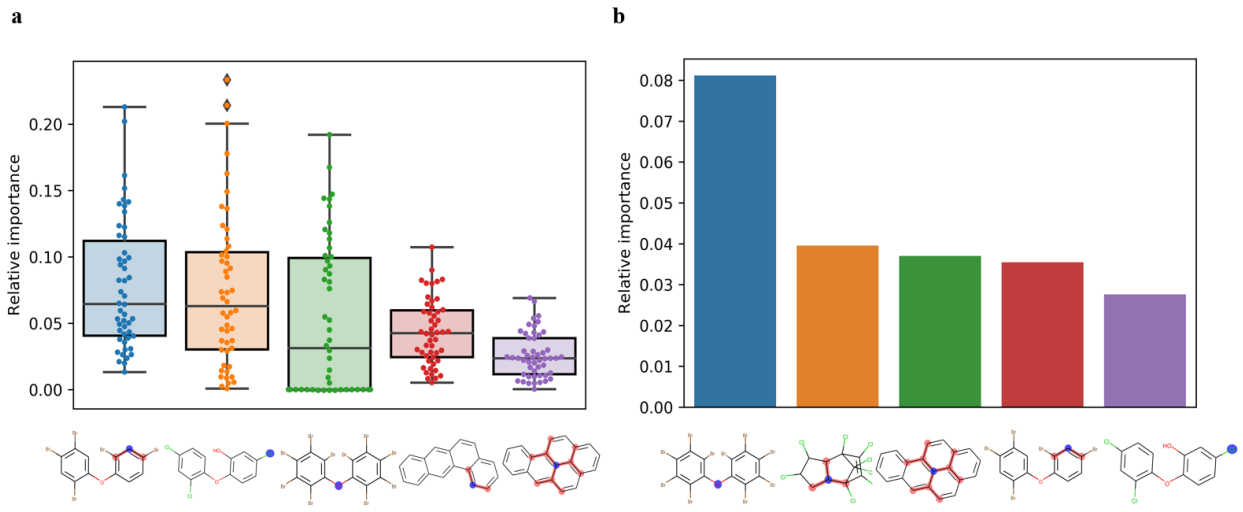

Fig. 4: Top 5 most important substructures with **a**, permutation importance; **b**, impurity importance in the dataset. Chemical selected randomly only to show key substructures.

Implications

RCF_{soil} predictions have been a challenge due to the complexity of compartments involved. The exploration of RCF_{water} to RCF_{soil} in field scale still lacks prediction power since the interactions between chemicals and plant roots are significantly affected by the composition

of the soil. Linear QSAR models cannot adequately describe the interactions in the ternary systems and can underestimate the risk of chemical' uptake from soil. Thus, it is essential to provide more accurate prediction for RCF_{soil} directly, which is needed urgently within the LCA framework to accurately characterize the human and ecotoxicological impacts of the fast-growing numbers of chemicals entering the environment. Deep learning and machine learning models with unbiased topological molecular fingerprints provide the opportunity to explore the relationship between molecular structures and RCF_{soil} covering both well-recognized competitive hydrophobic interactions among soil and plant compartments and less acknowledged interactions such as stereochemical chemical-bioreceptor interactions. This not only provides a more precise prediction model than traditional QSAR models for key ecotoxicity characterization factors in LCIA and other applications, but also helps to interrogate complicated chemical-soil-plant interactions. Furthermore, the refined model presented here and the key substructures it has identified can be used in the chemical development in the production of greener chemicals. Finally, based on the success of using machine learning models for RCF_{soil} prediction, future applications of the proposed model with larger datasets would be promising in dealing with broader aspects of life cycle analysis and environmental risk assessments.

Methods

Dataset collection

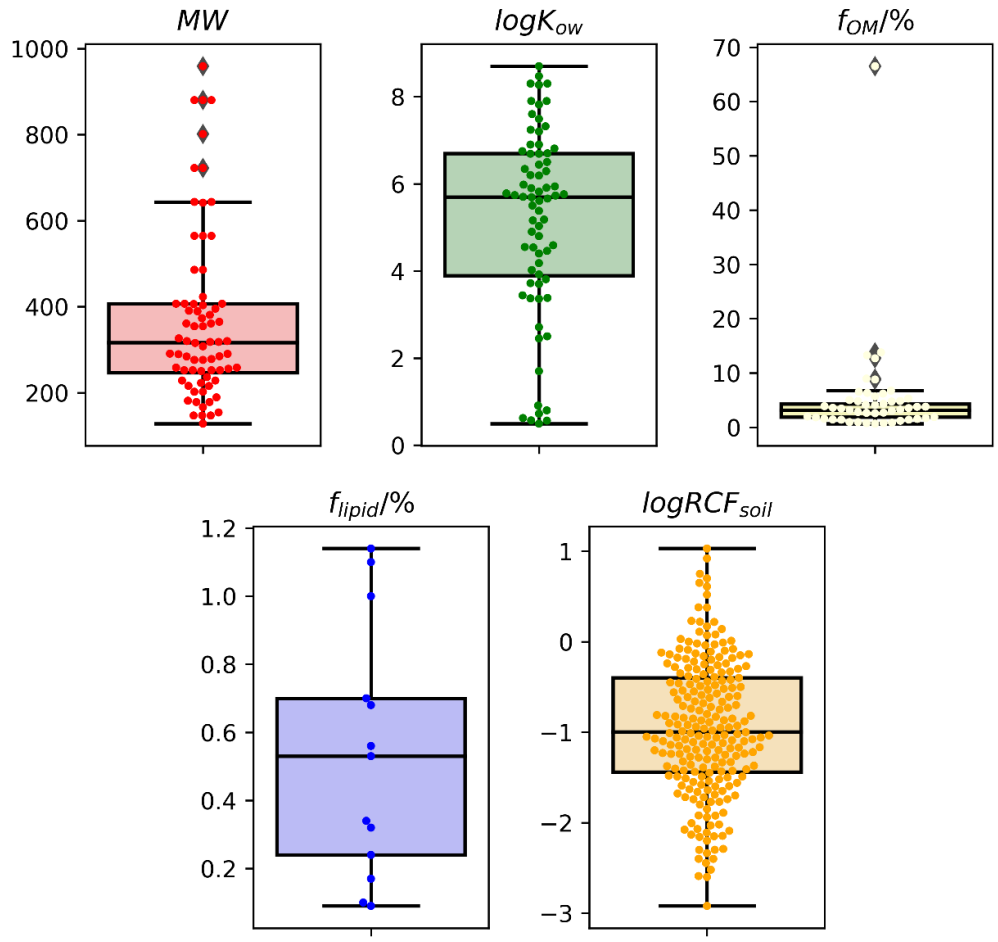

Fig. 5: Statistic analysis of the RCF_{soil} dataset. The dataset showed a large variation of chemicals with molecular weights ranging from 128 to 959 and $logRCF_{soil}$ ranging from -2.9 to 1.0.

A new plant uptake dataset that consists of 341 data points covering 72 chemicals was collected, together with their molecular structure specifications in the form of SMILES strings and related physio-chemical information such as $logK_{ow}$ and MW) (SI Table S3). For each data point, the corresponding f_{OM} and f_{lipid} were also included. The data set represents various plants, soil types and organic chemicals. Chemicals included polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), poly-brominated diphenyl ethers (PBDEs),

pesticides, pharmaceuticals, and many others. Plant species included wheat, carrot, radish, turnip, onion, spinach, celery, Chinese cabbage, pumpkin, maize, barely among others. The source data is comprised of a wide range of physicochemical properties with *MWs* ranging from 128 to 959 and $logK_{ow}$ from 0.50 to 8.70. The lipid content of the plant roots (f_{lipid}) covered a wide range of 0.10-6.7%. The soils also covered a broad range of physicochemical properties, with soil f_{OM} 0.69-66%. Statistical analysis of all included physicochemical properties were shown in Fig. 5. The root bioconcentration factors (RCF_{soil}) of chemicals from soils were defined as the ratio of plant root concentration (fresh weight) to soil concentration (dry weight). Only studies that reported soil f_{OM} and accessible plant f_{lipid} values were included in the dataset. This new dataset was included in SI for reference[25-45].

Molecular, plant and soil property representation

The predictive power of machine learning models relies not only on the algorithms but also on the quality of input data. The inputs to our machine learning models represented plant, soil, and molecular characteristics. Previous studies[7,46] suggested that the lipid fraction (f_{lipid}) in plant roots was the dominant factor controlling the uptake deviation observed among different plant species, indicating that using f_{lipid} would be a more general way of describing the related plant properties. Similarly, OM is the most important soil component that determines the sorption of organic chemicals, which is associated with the chemical's bioavailability for root uptake. Thus f_{OM} can be used to represent soil properties. The simplifications of the descriptors for plant root and soil properties allow developing models to focus on the effects of chemical structure on the uptake prediction. For molecular representations, two types of descriptors were compared: molecular physiochemical properties and molecular fingerprints. For molecular physiochemical properties, MW and $logK_{ow}$ were used. It should be noted here that although

multiple chemical properties could be included as molecular descriptors, previous studies have shown that MW and $logK_{ow}$ are the most important properties for root-water partitioning[12]. Another consideration of selecting physicochemical properties was that we tried to avoid those properties require expensive computation from molecule electronic structures, which may limit the model's applications in practice. For comparison, an empirical predictive LR model based on MW, $logK_{ow}$, f_{OM} and f_{lipid} was constructed to predict RCF_{soil}, as shown in Equation (1).

$$logRCF_{soil} = W_1 \times MW + W_2 \times logKow + W_3 \times f_{om} + W_4 \times f_{lipid} + b \quad (1)$$

Molecular representation with ECFP fingerprint

For molecular fingerprints, ECFP based on the Morgan Algorithm implemented in Python Rdkit package was employed[19]. ECFP is a circular fingerprint which represents the presence of particular substructures in a molecule. It was generated iteratively, starting from assigning an integer identifier for each atom. These initial atom identifiers formed the initial fingerprint set for the molecule. In the subsequent step, each atom's own identifier together with the identifiers of its immediate neighboring atoms were fed into a hash function to generate a new, single-integer identifier. These new identifiers were then added into the fingerprint set. This iteration is repeated for a prespecified number of times (2 in this study) and all duplicate identifiers in the set are removed. The remaining integer identifiers in the fingerprint set define the ECFP fingerprint. In this study, ECFP fingerprints with 1024 bits were used to avoid bit collision.

Gradient boosting regression tree

Gradient Boosting Regression Tree (GBRT) is a popular machine learning algorithm that has been widely used in many predictive models[47-49]. Gradient boosting is an additive model

which minimizes the loss function by adding weak learners. These weak learners are usually decision trees. The basic idea is that by adding a new tree into the model each time, the output of the existing sequence of trees will help improve the prediction, in this case RCF_{soil}, and decrease loss as shown in Fig. 1c. Our GBRT model utilized a mean squared error loss function defined as $Loss = \frac{1}{n}\sum^{n}\left(logRCF_{measured} - logRCF_{predicted}\right)^2$.

Model validation

Hyperparameters including max depth and number of estimators were carefully tuned through 5-fold cross validation as shown in Fig. 1d. A list of considered hyperparameter values can be found in SI Table S4. The whole dataset was first randomly split into five equal splits. Then each time, four of the five splits were used as training/validation set while the remaining split was left unseen by the model and used as test set to evaluate the model performance. Inside the training/validation set, 90% of the data was then used to train the model and the remaining 10% was used as the validation set. In other words, the model was trained on the training set, best hyperparameters were then chosen based on the performance on the validation set and finally model performance was evaluated with the chosen best hyperparameters on the test set. To fully test the model performance, each of the five sets was used as the test set once and eventually the model was validated by making predictions for each data point in the dataset. Finally, our model performance was evaluated by R^2 defined as $R^2(y, \hat{y}) = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, where $\hat{y}$ is the predicted value of i-th sample, y_i is the corresponding true value, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Data availability

The authors declare that all data supporting the findings of this study are available within the paper and its supplementary information files.

Code availability

All our models were implemented with Python Scikit-learn 0.24.1 package. ECFP were generated with Python RDkit 2020.3.3 package. The codes employed, as well as a notebook to reproduce our results, can be found in the public repository of this project (https://github.com/FengGmsu/RCF). Pseudocodes were also attached in SI.

References

1 Muir, D. C. & Howard, P. H. Are there other persistent organic pollutants? A challenge for environmental chemists. *Environ. Sci. Technol.* **40**, 7157-7166, doi:10.1021/es061677a (2006).

2 Miller, E. L., Nason, S. L., Karthikeyan, K. G. & Pedersen, J. A. Root uptake of pharmaceuticals and personal care product ingredients. *Environ. Sci. Technol.* **50**, 525-541, doi:10.1021/acs.est.5b01546 (2016).

3 Innovating the food value chain. *Nat. Sustain.* **3**, 1-1, doi:10.1038/s41893-020-0471-3 (2020).

4 Doucette, W. J., Shunthirasingham, C., Dettenmaier, E. M., Zaleski, R. T., Fantke, P. & Arnot, J. A. A review of measured bioaccumulation data on terrestrial plants for organic chemicals: metrics, variability, and the need for standardized measurement protocols. *Environ. Toxicol. Chem.* **37**, 21-33 (2018).

5 McKone, T. E. & Maddalena, R. L. Plant uptake of organic pollutants from soil: bioconcentration estimates based on models and experiments. *Environ. Toxicol. Chem.* **26**, 2494-2504, doi:10.1897/06-269.1 (2007).

6 Mamy, L., Patureau, D., Barriuso, E., Bedos, C., Bessac, F., Louchart, X., Martin-Laurent, F., Miege, C. & Benoit, P. Prediction of the fate of organic compounds in the environment from their molecular properties: A review. *Crit. Rev. Environ. Sci. Technol.* **45**, 1277-1377, doi:10.1080/10643389.2014.955627 (2015).

7 Trapp, S. Calibration of a plant uptake model with plant- and site-specific data for uptake of chlorinated organic compounds into radish. *Environ. Sci. Technol.* **49**, 395-402, doi:10.1021/es503437p (2014).

8 Torralba Sanchez, T. L., Liang, Y. & Di Toro, D. M. Estimating grass-soil bioconcentration of munitions compounds from molecular structure. *Environ. Sci. Technol.* **51**, 11205-11214, doi:10.1021/acs.est.7b02572 (2017).

9 Shen, Y., Li, H., Ryser, E. T. & Zhang, W. Comparing root concentration factors of antibiotics for lettuce (Lactuca sativa) measured in rhizosphere and bulk soils. *Chemosphere*, 127677 (2021).

10 Collins, C. D., Martin, I. & Fryer, M. Evaluation of models for predicting plant uptake of chemicals from soil. (2006).

<http://webarchive.nationalarchives.gov.uk/20140328084622/http://www.environment-agency.gov.uk/static/documents/Research/sc050021_2029764.pdf>.

11 Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547-555, doi:10.1038/s41586-018-0337-2 (2018).

12 Domingo-Almenara, X., Guijas, C., Billings, E., Montenegro-Burke, J. R., Uritboonthai, W., Aisporna, A. E., Chen, E., Benton, H. P. & Siuzdak, G. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nat. Commun.* **10**, 5811, doi:10.1038/s41467-019-13680-7 (2019).

13 Hou, P., Zhao, B., Jolliet, O., Zhu, J., Wang, P. & Xu, M. Rapid prediction of chemical ecotoxicity through genetic algorithm optimized neural network models. *ACS Sustain. Chem. Eng.* **8**, 12168-12176, doi:10.1021/acssuschemeng.0c03660 (2020).

14 Hou, P., Jolliet, O., Zhu, J. & Xu, M. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environ. Int.* **135**, 105393, doi:10.1016/j.envint.2019.105393 (2020).

15 Hino, M., Benami, E. & Brooks, N. Machine learning for environmental monitoring. *Nat. Sustain.* **1**, 583-588, doi:10.1038/s41893-018-0142-9 (2018).

16 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.* **54**, 7008-7018, doi:10.1021/acs.est.0c02526 (2020).

17 Sigmund, G., Gharasoo, M., Huffer, T. & Hofmann, T. Deep learning neural network approach for predicting the sorption of ionizable and polar organic pollutants to a wide range of carbonaceous materials. *Environ. Sci. Technol.* **54**, 4583-4591, doi:10.1021/acs.est.9b06287 (2020).

18 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.* **54**, 15385–15394, doi:10.1021/acs.est.0c03700 (2020).

19 Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742-754 (2010).

20 Li, Y., Chiou, C. T., Li, H. & Schnoor, J. L. Improved prediction of the bioconcentration factors of organic contaminants from soils into plant/crop roots by related physicochemical parameters. *Environ. Int.* **126**, 0160-4120 (2019).

21 Wang, S., Luo, C., Zhang, D., Wang, Y., Song, M., Yu, Z., Wang, Y. & Zhang, G. Reflection of stereoselectivity during the uptake and acropetal translocation of chiral PCBs in plants in the presence of copper. *Environ. Sci. Technol.* **51**, 13834-13841, doi:10.1021/acs.est.7b03350 (2017).

22 Garcia-Castro, M., Kremer, L., Reinkemeier, C. D., Unkelbach, C., Strohmann, C., Ziegler, S., Ostermann, C. & Kumar, K. De novo branching cascades for structural and functional diversity in small molecules. *Nat. Commun.* **6**, doi:10.1038/ncomms7516 (2015).

23 Carbonell, P., Carlsson, L. & Faulon, J.-L. Stereo signature molecular descriptor. *J. Chem. Inf. Model.* **53**, 887-897, doi:10.1021/ci300584r (2013).

24 Chiou, C. T., Sheng, G. & Manes, M. A partition-limited model for the plant uptake of organic contaminants from soil and water. *Environ. Sci. Technol.* **35**, 1437-1444, doi:10.1021/es0017561 (2001).

25 Kipopoulou, A., Manoli, E. & Samara, C. Bioconcentration of polycyclic aromatic hydrocarbons in vegetables grown in an industrial area. *Environ. Pollut.* **106**, 369-380 (1999).
26 Mikes, O., Cupr, P., Trapp, S. & Klanova, J. Uptake of polychlorinated biphenyls and organochlorine pesticides from soil and air into radishes (Raphanus sativus). *Environ. Pollut.* **157**, 488-496 (2009).
27 Beestman, G., Keeney, D. & Chesters, G. Dieldrin uptake by corn as affected by soil properties 1. *Agron. J.* **61**, 247-250 (1969).
28 Boxall, A. B., Johnson, P., Smith, E. J., Sinclair, C. J., Stutt, E. & Levy, L. S. Uptake of veterinary medicines from soils into plants. *J. Agric. Food Chem.* **54**, 2288-2297 (2006).
29 Cai, Q.-Y., Mo, C.-H., Wu, Q.-T. & Zeng, Q.-Y. Polycyclic aromatic hydrocarbons and phthalic acid esters in the soil–radish (Raphanus sativus) system with sewage sludge and compost application. *Bioresour. Technol.* **99**, 1830-1836 (2008).
30 Carter, L. J., Harris, E., Williams, M., Ryan, J. J., Kookana, R. S. & Boxall, A. B. Fate and uptake of pharmaceuticals in soil–plant systems. *J. Agric. Food Chem.* **62**, 816-825 (2014).
31 Gao, Y., Zhu, L. & Ling, W. Application of the partition-limited model for plant uptake of organic chemicals from soil and water. *Sci. Total Environ.* **336**, 171-182 (2005).
32 Huang, H., Zhang, S. & Christie, P. Plant uptake and dissipation of PBDEs in the soils of electronic waste recycling sites. *Environ. Pollut.* **159**, 238-243 (2011).
33 Jiang, L., Lin, J. L., Jia, L. X., Liu, Y., Pan, B., Yang, Y. & Lin, Y. Effects of two different organic amendments addition to soil on sorption–desorption, leaching, bioavailability of penconazole and the growth of wheat (Triticum aestivum L.). *J. Environ. Manage.* **167**, 130-138 (2016).
34 Macherius, A., Eggen, T., Lorenz, W. G., Reemtsma, T., Winkler, U. & Moeder, M. Uptake of galaxolide, tonalide, and triclosan by carrot, barley, and meadow fescue plants. *J. Agric. Food Chem.* **60**, 7785-7791 (2012).
35 Pannu, M. W., Toor, G. S., O'Connor, G. A. & Wilson, P. C. Toxicity and bioaccumulation of biosolids-borne triclosan in food crops. *Environ. Toxicol. Chem.* **31**, 2130-2137 (2012).
36 Prosser, R. S., Lissemore, L., Topp, E. & Sibley, P. K. Bioaccumulation of triclosan and triclocarban in plants grown in soils amended with municipal dewatered biosolids. *Environ. Toxicol. Chem.* **33**, 975-984 (2014).
37 Scheunert, I., Topp, E., Attar, A. & Korte, F. Uptake pathways of chlorobenzenes in plants and their correlation with n-octanol/water partition coefficients. *Ecotoxicol. Environ. Saf.* **27**, 90-104 (1994).
38 Tao, Y., Zhang, S., Zhu, Y.-g. & Christie, P. Uptake and acropetal translocation of polycyclic aromatic hydrocarbons by wheat (Triticum aestivum L.) grown in field-contaminated soil. *Environ. Sci. Technol.* **43**, 3556-3560 (2009).
39 Trapp, S., Matthies, M., Scheunert, I. & Topp, E. M. Modeling the bioconcentration of organic chemicals in plants. *Environ. Sci. Technol.* **24**, 1246-1252 (1990).
40 Wu, C., Spongberg, A. L., Witter, J. D. & Sridhar, B. M. Transfer of wastewater associated pharmaceuticals and personal care products to crop plants from biosolids treated soil. *Ecotoxicol. Environ. Saf.* **85**, 104-109 (2012).

41 Zhu, H., Sun, H., Zhang, Y., Xu, J., Li, B. & Zhou, Q. Uptake pathway, translocation, and isomerization of hexabromocyclododecane diastereoisomers by wheat in closed chambers. *Environ. Sci. Technol.* **50**, 2652-2659 (2016).
42 Wang, F., Li, X., Yu, S., He, S., Cao, D., Yao, S., Fang, H. & Yu, Y. Chemical factors affecting uptake and translocation of six pesticides in soil by maize (Zea mays L.). *J. Hazard. Mater.*, 124269, doi:10.1016/j.jhazmat.2020.124269 (2020).
43 Liu, Y., Ma, L. Y., Lu, Y. C., Jiang, S. S., Wu, H. J. & Yang, H. Comprehensive analysis of degradation and accumulation of ametryn in soils and in wheat, maize, ryegrass and alfalfa plants. *Ecotoxicol. Environ. Saf.* **140**, 264-270, doi:10.1016/j.ecoenv.2017.02.053 (2017).
44 Harris, C. R. & Sans, W. Absorption of organochlorine insecticide residues from agricultural soils by root crops. *J. Agric. Food Chem.* **15**, 861-863 (1967).
45 Zhang, J., Zhao, W., Pan, J., Qiu, L. & Zhu, Y. Tissue-dependent distribution and accumulation of chlorobenzenes by vegetables in urban area. *Environ. Int.* **31**, 855-860 (2005).
46 Hung, W.-N., Chiou, C. T. & Lin, T.-F. Lipid–water partition coefficients and correlations with uptakes by algae of organic compounds. *J. Hazard. Mater.* **279**, 197-202, doi:10.1016/j.jhazmat.2014.06.071 (2014).
47 Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. & Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56-67, doi:10.1038/s42256-019-0138-9 (2020).
48 Isayev, O., Oses, C., Toher, C., Gossett, E., Curtarolo, S. & Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, doi:10.1038/ncomms15679 (2017).
49 Adeogba, E., Barty, P., O'Dwyer, E. & Guo, M. Waste-to-resource transformation: gradient boosting modeling for organic fraction municipal solid waste projection. *ACS Sustain. Chem. Eng.* **7**, 10460-10466, doi:10.1021/acssuschemeng.9b00821 (2019).

Acknowledgements

The work was supported by the National Key Research and Development Program of China (2016YFD0800403, 2019YFC1604503 and 2020YFC1806801).

Author Contributions

F.G. and Y. L. conceived the idea. Y.L. collected data. F.G., C.L. and Y. L. developed the machine learning models and wrote the manuscript. Y.L., Y.S., B.S., and H.L. contributed to the interpretation of the results, revised, and edited the manuscript.

Supplementary information

Supplementary Results, Figs. S1-2, and Tables S1-4, Table S2 was attached in a separate spreadsheet. Pseudocode of the machine learning method was also included.

Competing interests

The authors declare no competing interests.

Figures

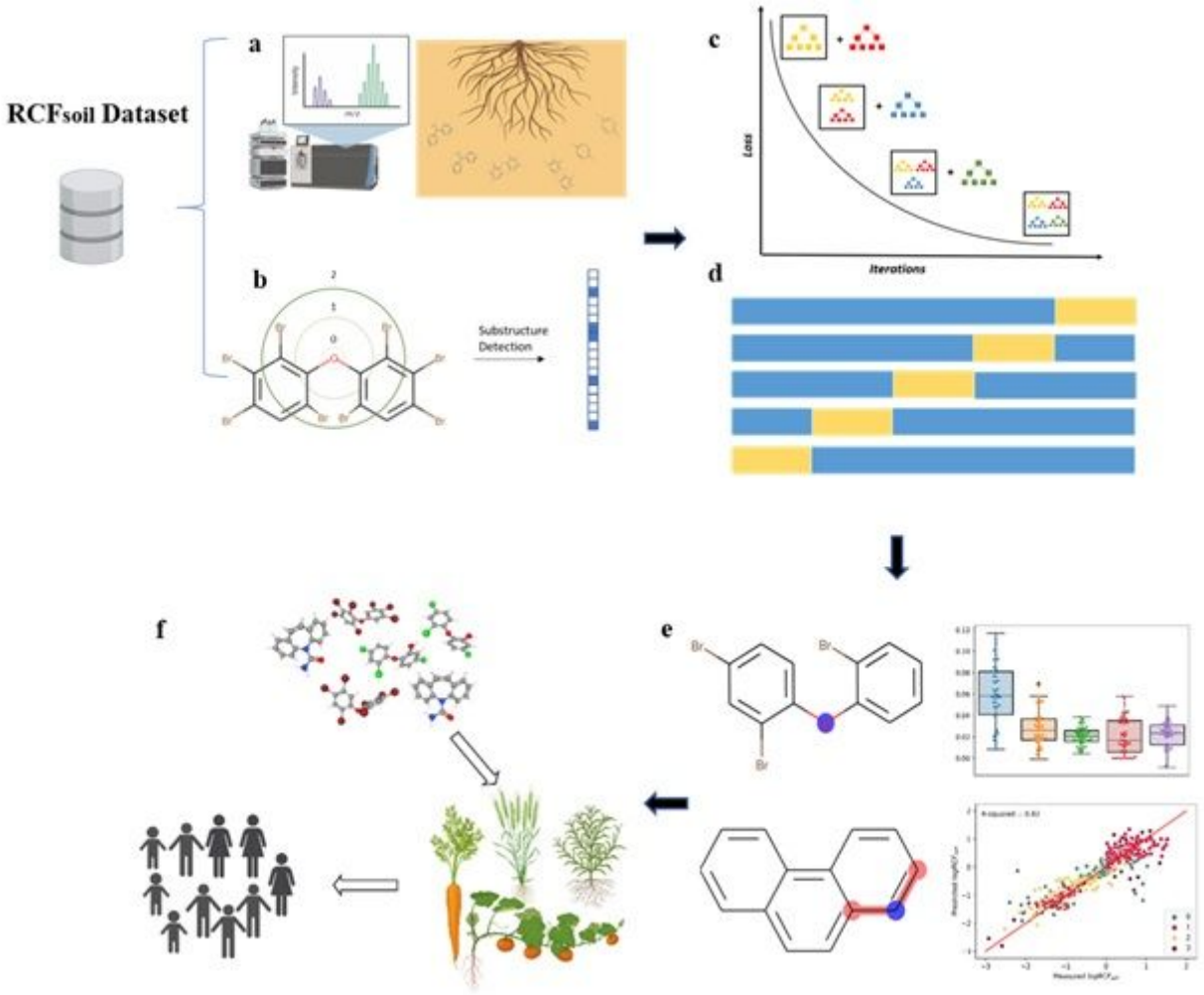

Figure 1

Schematic representation for building machine learning model for RCF_soil prediction and implications: a, Data collected from plant uptake studies. b, ECFP used for molecular structures representation. c, Gradient boosting regression tree model trained on the dataset. d, 5-fold cross validation for parameter tuning. e, RCF_soil prediction and model interpretation. f, Implication of model results in LCIA.

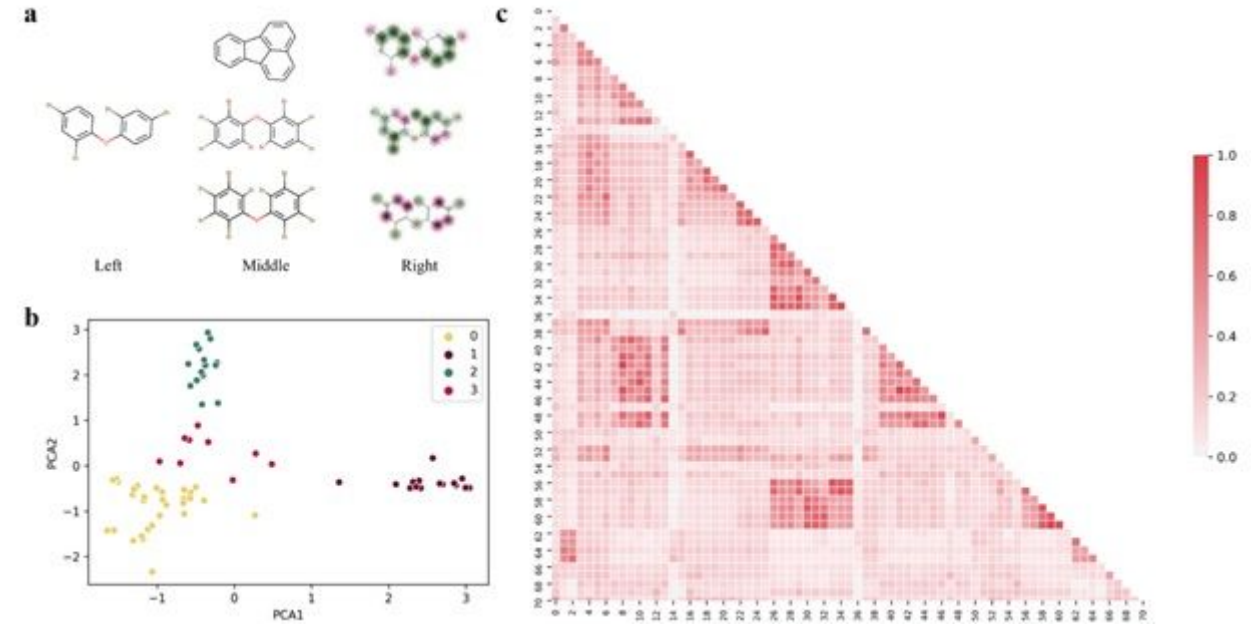

Figure 2

a, Comparison of 2,2',4,4'-tetrabromodiphenyl ether (BDE-47, left) with three molecules (middle) from top to bottom: fluoranthene, 2,2',3,3',4,4',6,6'-octabromodiphenyl ether (BDE-197), decabromodiphenyl ether (BDE-209). Comparison results (right) with dark green showing similar substructures while dark red showing different substructures. b, Clustering of chemicals based on ECFP fingerprints. c, Similarity comparison of chemicals in the dataset (darker red shows higher similarities).

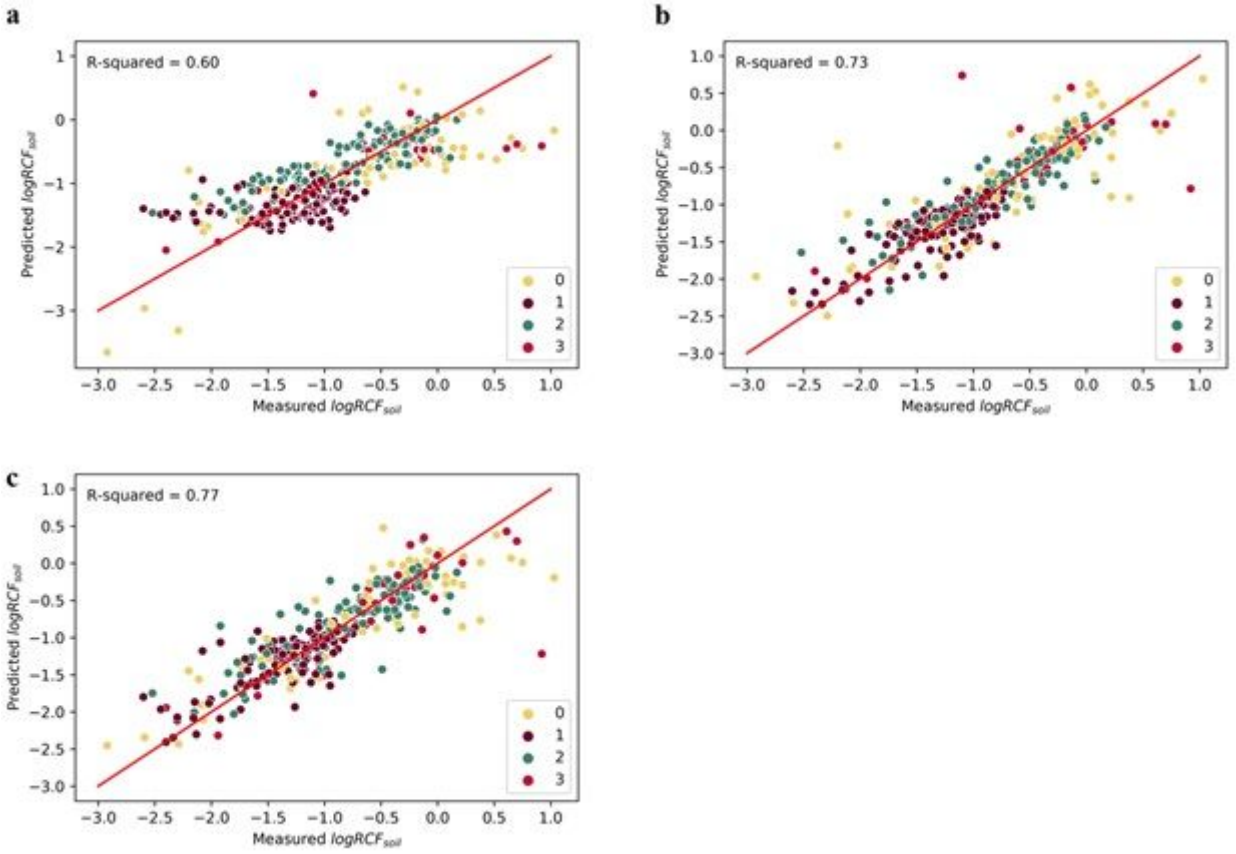

Figure 3

Prediction of logRCF_soil with a, LR model and molecular physicochemical property descriptors; b, GBRT model and molecular physicochemical property descriptors; and c, ECFP-GBRT model. Different colors represent different chemical groups from clustering.

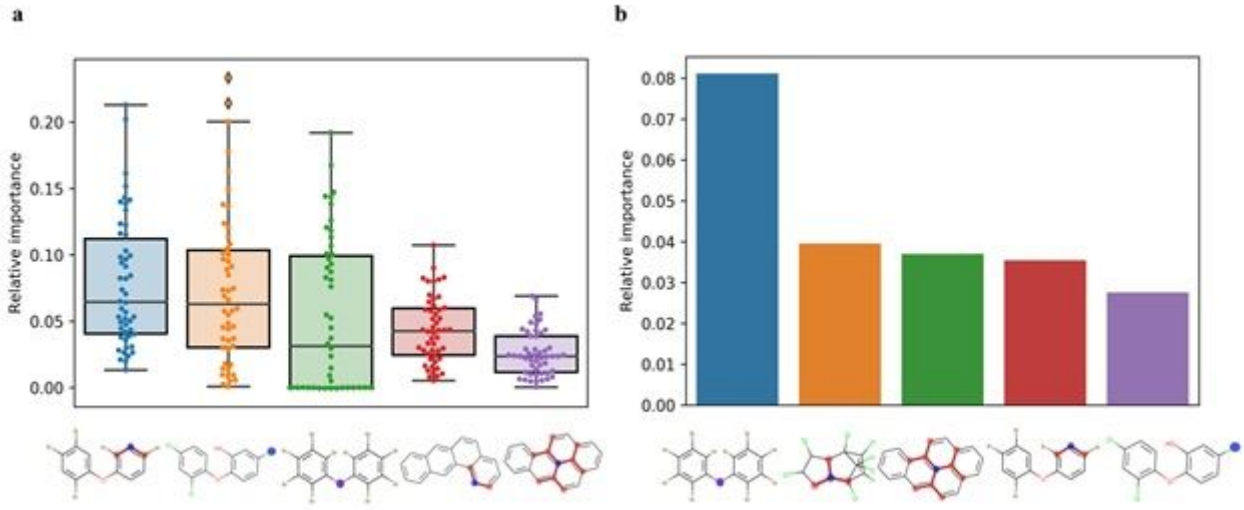

Figure 4

Top 5 most important substructures with a, permutation importance; b, impurity importance in the dataset. Chemical selected randomly only to show key substructures.

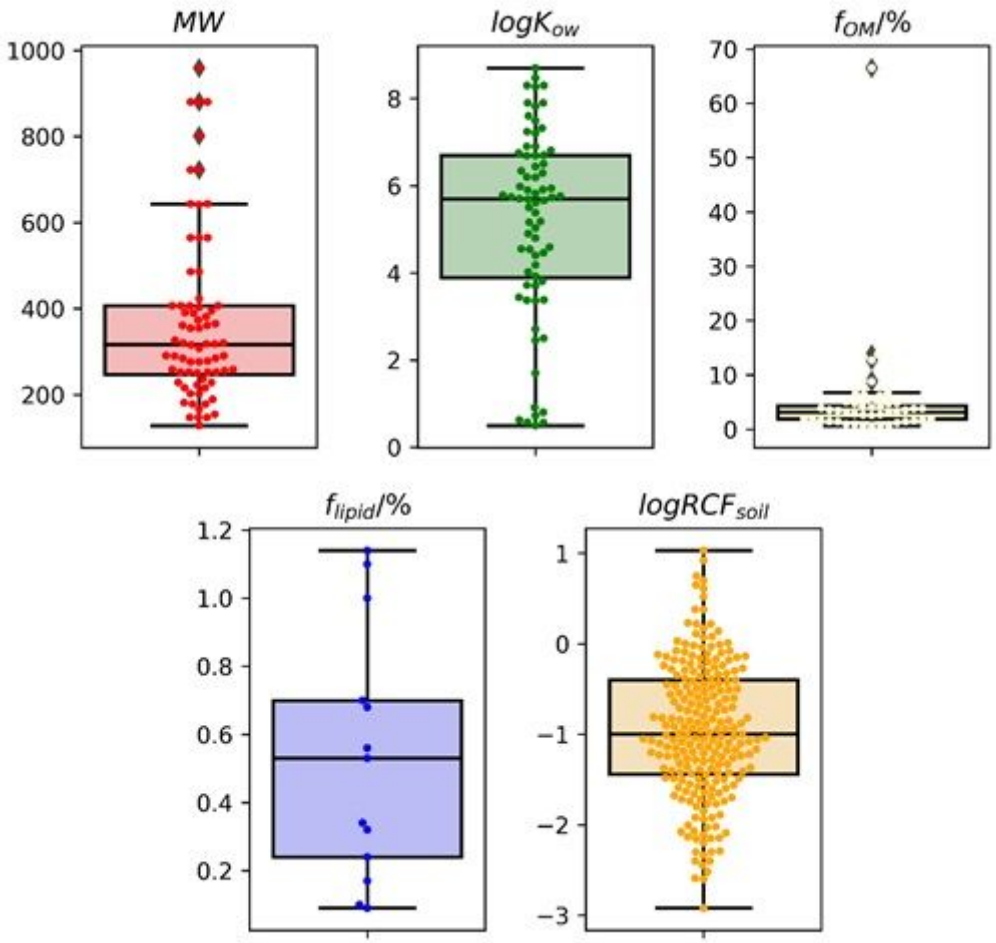

Figure 5

Statistic analysis of the RCF_soil dataset. The dataset showed a large variation of chemicals with molecular weights ranging from 128 to 959 and logRCF_soil ranging from -2.9 to 1.0.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DirectpredictionofRCFsoilMLSItableS2heatmapdataframe.pdf
- DirectpredictionofRCFsoilMLSI.docx
- NATSUSTAIN21028965SI.pdf