Environmental Science Processes & Impacts

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5em00532a

Molecular connectivity indices and soil properties to predict the sorption of per- and poly-fluoroalkyl substances

Paulina Alulema-Pullupaxi, [ac] Fatih Evrendilek, [bc] Dilara Hatinoglu, [ac] Simin Moavenzadeh Ghaznavi,[c] Kenneth Mensah, [ac] Manisha Choudhary, [ce] Sonora Ortiz [d] and Onur Apul *[ac]

This study presents a modeling approach to predict the soil-water partitioning coefficient (K_d, L kg^{-1}) for per- and poly-fluoroalkyl substances (PFASs) as a function of their molecular connectivity indices (MCIs) and soil properties (soil organic carbon, SOC, %, and cation exchange capacity, CEC, cmol kg^{-1}). The modeling framework involved compiling data, developing models, and evaluating model performance *via* interpretation, external validation, and scenario analyses. Two datasets consisting of simple and valence MCIs per PFAS were used: (i) carboxylic-PFCA dataset (N = 327) had only carboxylic compounds (C4–C12) and (ii) PFAS-full dataset (N = 699) entailed carboxylic acids (C4–C12), sulfonic acids (C4–C10) and fluorotelomers (C4–C8). Our multi-criteria approach revealed that the seventh-order valence path (VP-7) related to polarizability and molecular size and the third-order simple path (SP-3) related to molecular size and chain structure emerged as key predictors for the carboxylic-PFAS and PFAS-full datasets, respectively. Elastic net-regularized linear regression (MLR_{EN}) and artificial neural networks (ANNs) demonstrated that MCIs improved the predictive accuracy. For the PFAS-full dataset, six-predictor models (MCIs + soil properties) yielded a high predictive accuracy (R_{pred}^2 = 83.7–84.9%); however, a three-predictor MLR_{EN} model (SP-3, SOC, and CEC; R_{pred}^2 = 77.9%) achieved the highest external generalization (R_{ext}^2 = 52.4%). SP-3 accounted for the largest share of predictive power (68–95%), dominating the model performance (94–97%). Scenario analyses revealed that while deterministic predictions remained stable, probabilistic modeling is crucial for capturing the rare but impactful extremes. Overall, our study highlights the practical advantage of MCIs as versatile and scalable tools for predicting the adsorption of diverse PFAS, including short-chain, partially fluorinated, and less commonly studied PFASs. In the long term, this tool can provide data for preliminary, rapid, site-specific risk assessment for PFAS-impacted sites.

Received 10th July 2025
Accepted 30th March 2026

DOI: 10.1039/d5em00532a

rsc.li/espi

Environmental significance

Per- and poly-fluoroalkyl substances (PFAS) are persistent, mobile, and toxic environmental contaminants that threaten soil and groundwater resources. Understanding and predicting how PFAS interact with soils is critical for managing their risks. However, laboratory testing of PFAS sorption is time- and resource-intensive, particularly for emerging or understudied compounds. This study demonstrates a novel and scalable modeling framework that integrates molecular connectivity indices (MCIs) with soil properties to predict the PFAS sorption behavior. By capturing their key molecular features such as size and branching, this model enables rapid estimation of PFAS mobility across a wide range of compounds and soils. This approach can serve as a screening tool for site-specific risk assessment and help prioritize remediation efforts. The integration of deterministic and probabilistic analyses also enhances environmental decision-making by identifying potential high-risk scenarios.

1 Introduction

Soils can act as the sink and secondary source of PFAS, especially at sites impacted by wastewater biosolid application, industrial discharges, or firefighting foam spills.[1–3] The remarkable persistence and widespread distribution of PFAS have led to significant environmental pollution, underscoring the need for robust predictive models to assess their

[a]Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, 16802, USA. E-mail: oga5061@psu.edu
[b]University of Maine Cooperative Extension, Orono, ME 04469, USA
[c]Department of Civil and Environmental Engineering, University of Maine, Orono, ME, 04469, USA
[d]Department of Ecology and Environmental Sciences, University of Maine, Orono, ME, 04469, USA
[e]Department of Biological and Agricultural Engineering, Kansas State University, Manhattan, KS, 66506, USA

environmental fate and transport.[4–6] The accurate prediction of PFAS sorption to soil, commonly quantified by the soil-water partition coefficient (K_d, L kg^{-1}), is crucial for assessing the environmental risk and designing effective remediation strategies. However, PFAS sorption is a complex process governed by multiple mechanisms (*e.g.*, hydrophobic interactions, electrostatic interactions, and ligand exchange) that depend on both the specific PFAS molecular structure and key soil properties, such as soil organic carbon (SOC), cation exchange capacity (CEC), soil pH, and clay mineralogy.[7–10] Given all these factors, conducting laboratory experiments to determine the K_d is highly labor- and resource-intensive.[9,11] Therefore, the PFAS molecular structure and edaphic drivers can be effectively incorporated into predictive modeling approaches for risk assessment and environmental management.[8,9,12]

Existing predictive models such as quantitative structure–activity relationship (QSAR), quantitative structure–property relationship (QSPR), and linear-solvation energy relationship (LSER) have been reported to predict parameters that govern the PFAS environmental behavior.[13–17] These models typically use physicochemical descriptors, molecular properties (*i.e.*, molar volume, molecular weight, fluorine number, carbon number, and carbon number in tail), and solvatochromic Abraham descriptors as explanatory variables and primarily target long-chain PFAS (either with carboxylic or sulfonic functional groups). However, these models often fail to represent the broader chemical diversity of PFAS, particularly emerging short-chain compounds, branched compounds or partially fluorinated precursors like fluorotelomer sulfonates (FTS), which can degrade into short-chain regulated compounds.[18,19] In particular, LSER models, although mechanistically informative, are limited to neutral compounds, which restricts their use under diverse environmental conditions for ionizable PFAS. For this, our previous study has investigated the adjustment of solvatochromic predictors for carboxylic PFAS.[13,20] To advance this approach, in accordance with the rapidly developing PFAS literature, we now explore molecular connectivity indices (MCIs) as a versatile and chemically inclusive alternative for a comprehensive predictive tool.

MCIs are topological descriptors derived from the molecular graph, quantifying aspects of molecular size, branching, and derived electronic properties based on atom connectivity.[21] Unlike descriptors requiring specific functional group information or 3D conformation, MCIs encode structural information inherent in the bonding topology, potentially offering broader applicability across diverse PFAS structures, including those lacking traditional functional groups or with complex branching.[22] These descriptors have proven successful in predicting octanol–water (K_{ow}) and octanol–air (K_{oa}) partition coefficients,[23,24] bioconcentration ratios,[25,26] and distribution coefficients for the sorption of aromatic compounds by carbon-based adsorbents and soils,[27–29] where zero-, first-, and third-order simple and valence indices were found to be the best topological predictors.[23,25–28] However, their systematic application and evaluation for the modeling partitioning of PFAS sorption in soil (K_d) remain unexplored, representing a critical knowledge gap.

Therefore, our study is the first systematic development and validation of MCIs for predicting the PFAS sorption in soils in combination with soil physicochemical attributes. To achieve this goal, the research was designed with four specific objectives. First, we provide a baseline for validating the MCIs framework within a chemically homogeneous subset by comparing the predictive performance of MCIs with Abraham-solvatochromic descriptors to predict $\log K_d$ of perfluoroalkyl carboxylic acids (PFCAs) before extending the analysis to a broad structurally diverse set of PFAS. Second, we develop and evaluate the predictive and generalization abilities of linear and machine-learning models to predict $\log K_d$ for multiple PFAS subclasses (*i.e.*, perfluoroalkyl carboxylic acids – PFCAs, perfluoroalkyl sulfonic acids – PFSAs, and fluorotelomer sulfonates – FTS). Third, we characterize the predictor importance and interaction effects of molecular and soil predictors using Monte Carlo simulations to elucidate key factors influencing PFAS sorption. Finally, we conduct scenario analyses to identify the conditions that maximize PFAS sorption ($\log K_d$) *via* composite desirability function (D) and evaluate model uncertainty and robustness under boundary minima *via* Monte Carlo simulations.

2 Data collection, model development and model evaluation

Fig. 1 outlines the systematic workflow adopted in this study, which comprises three stages: (1) data collection, (2) model development, and (3) model evaluation. This workflow integrates the three datasets used for model development and evaluation.

2.1. Data collection

Three distinct datasets compiled from the peer-reviewed literature and used for this study are shown in Fig. 1. For all datasets, the response variable was literature-reported soil–water partition coefficient ($\log K_d$) of the group of PFAS included in each dataset, while explanatory variables included both PFAS molecular descriptors (*i.e.*, molecular connectivity indices (MCIs) and Abraham descriptors) and soil physicochemical properties (*i.e.*, soil organic carbon – SOC, %; cationic exchange capacity – CEC, cmol kg^{-1}; and soil pH).

The first dataset ($N = 327$), called carboxylic–PFAS dataset, contained nine carboxylic PFAS (C4–C12), and was used to evaluate the predictive capacity of MCIs by comparing them against Abraham descriptors. The second dataset ($N = 699$), called PFAS-full dataset, extended the chemical coverage that include 9 perfluorocarboxylic acids – PFCAs (C4–C12), 7 acid perfluorosulfonic acids – PFSAs (C4–C10), and 3 fluorotelomer sulfonates – FTS (C4–C8), and it was used for developing and comparing MCIs-based predictive models for PFAS. Both datasets were solely used for model development and internal validation, each randomly split into 75% training and 25% validation subsets.

Molecular descriptor calculation began with retrieving the simplified molecular input line entry system (SMILES) strings

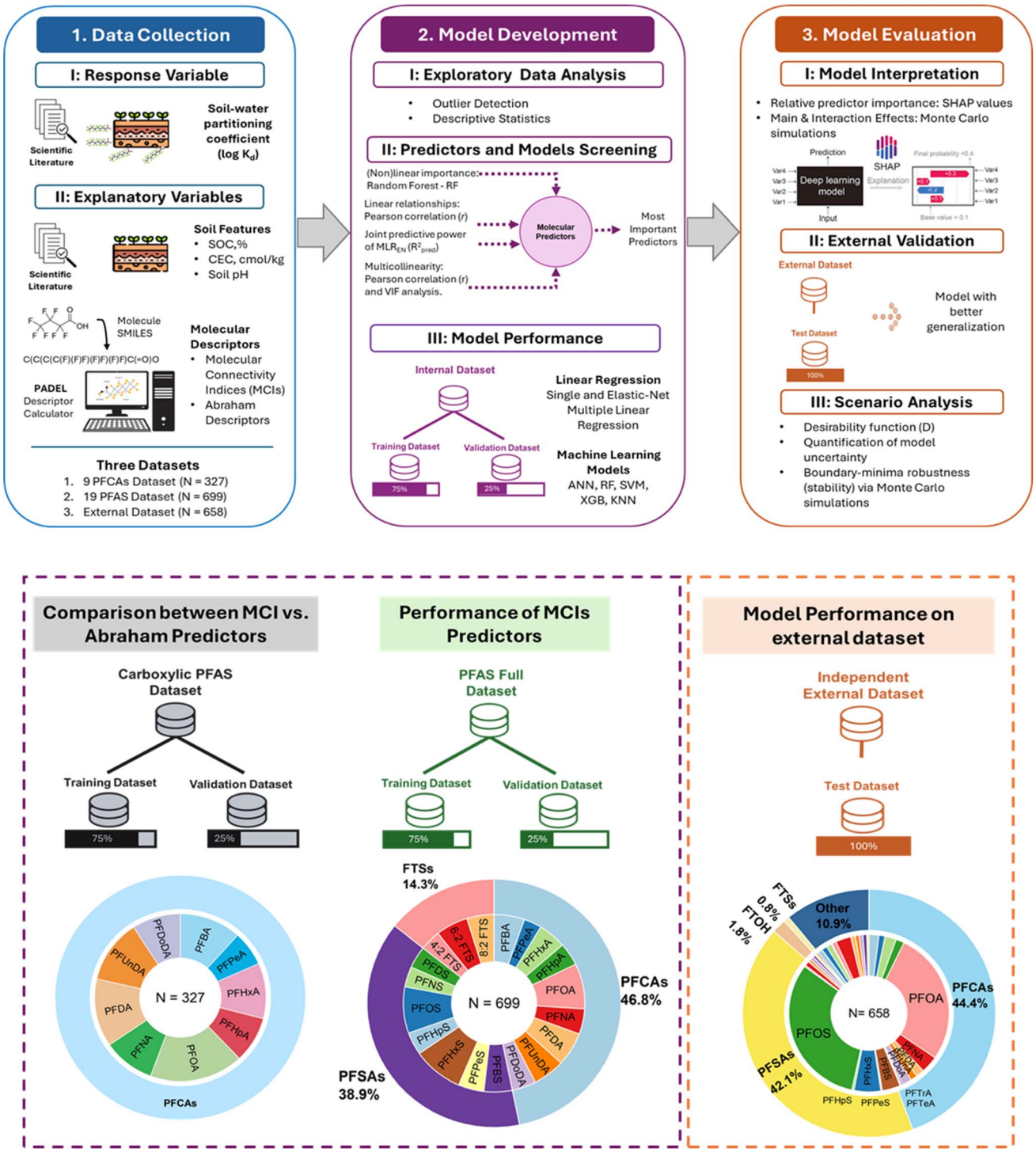

Fig. 1 Workflow followed in this study divided into three stages: (1) data collection, (2) model development, and (3) model evaluation. Dataset splitting of each dataset was used in the internal training and validation of predictive models and independent external validation of the best-performance models. Soil organic carbon (SOC), cation exchange capacity (CEC), sample size (*N*), elastic net-regularized regression (MLR_{EN}), variance inflation factor (VIF), artificial neural networks (ANN), random forest (RF), support vector machine (SVM), extreme gradient boosting (XGB), K nearest neighbor (KNN), and Shapley additive explanations (SHAP).

for each PFAS molecule in both neutral and ionic forms from the PubChem database.[30] Using the ChemDes platform (https://www.scbdd.com/chemdes) and the PaDEL Descriptor Calculator,[31] SMILES were transformed into 46 MCIs in acidic and ionic forms of each compund. This study included the following two classes: (1) simple indices that encode sigma-bonding patterns from structural formula and (2) valence indices that incorporate sigma, pi, and lone-pair electrons, thus

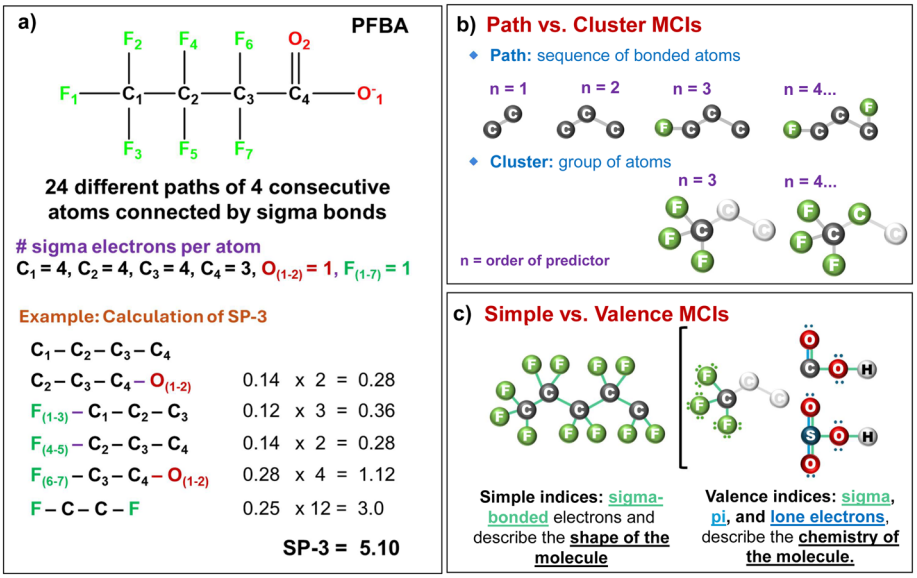

Fig. 2 Schematic of the molecular connectivity indices (MCIs) for PFAS. (a) Illustration of the calculation of third-order simple path (SP-3) molecular connectivity index for PFBA; (b) index order and fragment types of PFAS molecules with arrangement of atoms and bonds into path or cluster; and (c) simple and valence index configurations of PFAS: simple indices encode sigma bonds, while valence indices incorporate sigma, pi, and lone-pair electrons.

more detailed electronic information. Index order (n = 0–7, zeroth to seventh order) and fragment type (*e.g.*, simple or valence path, SP or VP; simple or valence cluster, SC or VC) define the structural scope of each MCIs, as described in Text S1. Fig. 2 illustrates the derivation and structural interpretation of MCIs, using perfluorobutanoic acid (PFBA) as an example. These descriptors capture how atomic connectivity influences molecular behavior, providing a structural basis for linking PFAS chemistry to sorption processes in soils. On the other hand, Abraham descriptors were calculated for both neutral and ionic forms following the methodology report by Hatinoglu *et al.* (2023),[13] with a total of five descriptors at neutral form and six descriptors at ionic form (′) of each compound: excess molar refraction (E, E'), dipolarity/polarizability (S, S'), hydrogen bond acidity (A, A'), hydrogen bond basicity (B, B'), and molar volume (V, V), and an additional descriptor (J^-) to distinguish ionized from neutral species (Table S4). These descriptors are widely used in LSER models to quantify solute–solvent interactions. Both sets of descriptors (MCIs and Abraham descriptors) were corrected to reflect PFAS dissociation states under experimental conditions, as summarized in Text S1.

The experimental K_d values and corresponding soil properties were compiled from literature, specifically from peer-reviewed articles that are reporting experimentally derived sorption isotherms, following the methodology reported in our recent publication.[32] The sorption isotherm slope, K_d, was derived from linear fits, with consideration of test conditions such as SOC (%), CEC (cmol kg^{-1}), pH, and the structural characteristics of PFAS compounds, including short and long C-chains and functional groups. When adsorption was non-linear and K_d was not directly reported, K_d was derived from the initial linear (Henry) region of the isotherm, corresponding to $C_e \approx 10^{-5}$ mg L^{-1}, where sorption is proportional to solute concentration and independent of site saturation. The concentration of a compound adsorbed to the soil matrix (C_s, mg kg^{-1}) and its concentration in the aqueous phase (C_w, mg L^{-1}) from this region were used to calculate K_d. For datasets reporting K_{oc} only (organic carbon normalized distribution coefficient), K_d was back calculated using the reported f_{oc} (fraction of the solid that is organic carbon). All values were standardized to mg kg^{-1} and mg L^{-1} for consistency.[32] The compiled datasets encompassed a wide range of soil matrices, including uncontaminated reference soils,[9,33–43] AFFF-contaminated soils,[11,44–46] and pure clay minerals[47,48] (*e.g.*, montmorillonite, kaolinite, and illite), collected across the U.S., Canada, Sweden, China, and South Africa, capturing global variability in soil composition and contamination sources (Table S2).

The third dataset (N = 658), called independent external dataset, was partially compiled from a recent published study[49] and contained log K_d (L kg^{-1}), SOC (%), and CEC (cmol kg^{-1}) data. To evaluate model generalization, it was subdivided into soil-only and combined soil & sediments groups, as presented in Fig. 3. The soil-only subset consisted of 31 PFAS across multiple types: 11 PFCAs (C4–C14), 6 PFSAs (C4–C10), 2 FTS (C4–C6), 4 fluorotelomer alcohols – FTOHs (C4–C10), 2 zwitterionic PFAS, 3 perfluoro phosphonic acids – PFPAs (C6–C10); plus 8 : 2 chlorinated polyfluoroalkyl ether sulfonic acid – 8 : 2Cl-PFESA, trifluoroacetic acid – TFA, and perfluorooctaneamido ammonium – PFOAAmS, while the soil & sediments subset included the same 31 PFAS types plus four additional compounds (*i.e.*, *N*-methyl perfluorooctanesulfonamido acetic acid – *N*-MeFOSAA, perfluorodecanoic sulfonic acid – PFDS, *N*-ethyl perfluorooctanesulfonamido acetic acid – *N*-EtFOSAA, and perfluorooctane sulfonamide – PFOSA). Each

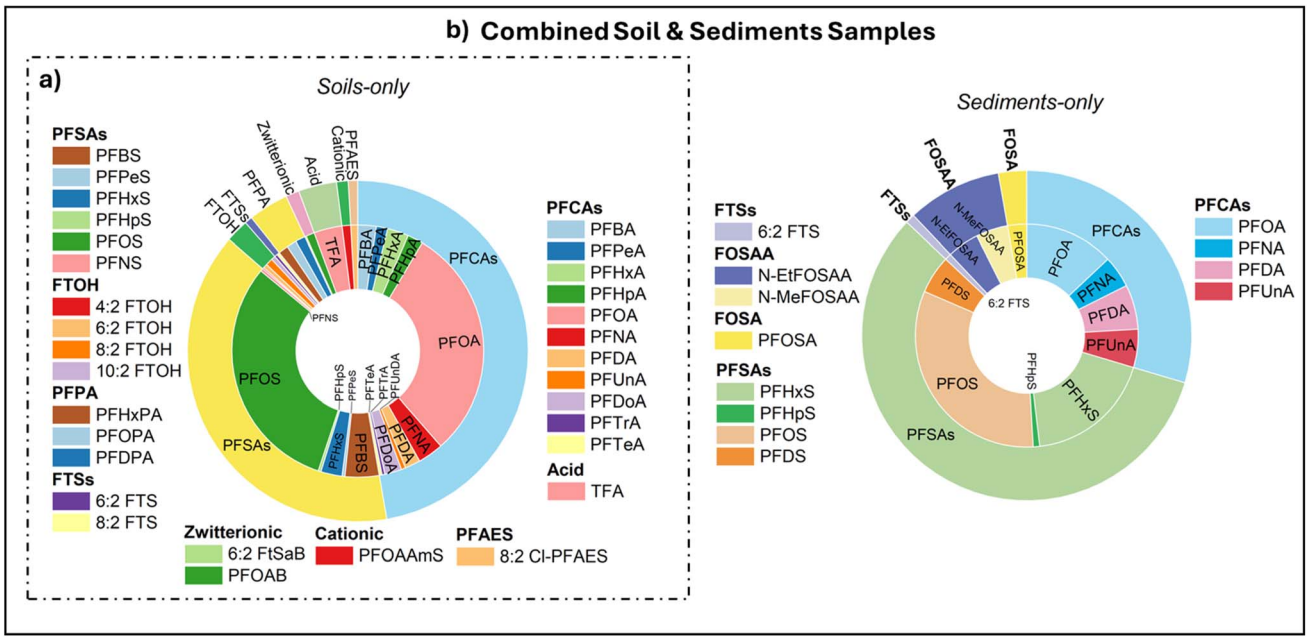

Fig. 3 PFAS subclasses included in the independent external dataset subdivided by their environmental matrix type: (a) soil-only and (b) combined soil and sediment samples.

subset was exclusively used for external validation to assess generalization capability across environmental conditions, ensuring full independence from training data and preventing data leakage. No normalization, scaling or transformation was applied prior to modeling to preserve original data distributions.

2.2. Model development

The predictive modeling framework for estimating $\log K_d$ was developed following three integrated stages: (i) exploratory data analysis; (ii) predictor and model screening; and (iii) internal model predictive performance using carboxylic-PFAS and PFAS-full datasets. The first stage included the calculation of descriptive statistics (Tables S5, S6 and Fig. S1–S3), and identifying potential outliers could distort observed PFAS–soil relationships. Outlier detection followed a consistent procedure across carboxylic-PFAS and PFAS-full datasets using a statistical criterion on the interquartile range (IQR).[50] Specifically, data points lying outside the range defined by the 10th and 90th percentiles $\pm 3 \times$ IQR were identified and removed. Outliers removed included one CEC value from the carboxylic-PFAS dataset, 51 SOC and two CEC outliers from the PFAS-full dataset. For the independent external dataset, we applied our own outlier detection, as stated above, to ensure consistency with our preprocessing pipeline, even though the original source may have its own criteria. A total of 30 SOC outliers were removed from this dataset.

Second stage focused on identifying the most relevant MCIs and soil predictors while preventing model overfitting. A non-formal feature screening process was adopted, integrating multiple-criteria analyses to balance statistical robustness with environmental interpretability. Non-linear importance was assessed using the Random Forest (RF) algorithm to evaluate the relative contribution of each molecular connectivity index to $\log K_d$.[50] Linear relationships were examined through Pearson's correlation coefficient (r) to evaluate both the strength of association between each predictor and $\log K_d$ and the degree of multicollinearity among predictors.[51] Joint predictive power was preliminarily evaluated using elastic net-regularized multiple linear regression (MLR_{EN}; $\alpha = 0.99$; number of grid points = 150; grid scale = square root; and minimum penalty fraction = 0.001) to asses the combined predictive contribution of individual MCIs when integrated with soil properties.[52] Multicollinearity was controlled by retaining only predictors with a variance inflation factor (VIF) ≤ 12. The final subset was selected based on a multi-criteria approach combining RF importance, strong correlation with $\log K_d$, strong predictive performance in preliminary MLR_{EN} models, and low multicollinearity (VIF ≤ 12), reflecting variables that were both statistically stable and mechanistically meaningful in describing PFAS–soil interactions. This exploratory process prioritized interpretability and reproducibility over exhaustive optimization; hence, no hyperparameter tuning was conducted. The retained predictors are summarized in Table S7.

The third stage involved model training and validation phase. Both the carboxylic-PFAS and PFAS-full datasets were randomly partitioned into 75% training and 25% validation subsets to ensure the balanced representation of compound types. The carboxylic-PFAS dataset was analyzed separately to provide a chemically controlled baseline, enabling direct comparison between the new MCIs framework and established Abraham solvatochromic descriptors that have been only developed and validated for carboxylic PFAS. This comparison validated the MCIs approach within a homogeneous compound

group before extending it to the more structurally diverse PFAS-full dataset, which includes carboxylic, sulfonic, and fluorotelomer sulfonates. Subsequently, linear and machine learning models were developed using the PFAS-full dataset to assess the predictive capabilities of MCIs across compound classes. Simple linear regression (SLR) models were trained using selected MCIs (*i.e.*, VP-1, VP-4, VC-3, VC-4, VC-6, VPC-4, SP-3, SP-4, and SPC-4), as described in Table S9. MLR_{EN} was subsequently fitted using either the same predictors as SLR or an extended set including ASP-1, ASP-0, AVP-0 and soil properties, as shown in Table S10a. Additional machine learning algorithms including artificial neural networks (ANN), random forest (RF), support vector machine (SVM), extreme gradient boosting (XGB), and K nearest neighbors (KNN) were trained using multiple MCIs (*i.e.*, SP-3, ASP-1, AVP-0, and ASP-0) and soil properties (Table S10a). All models were implemented using the default hyperparameter settings of their respective libraries, and no hyperparameter optimization or cross-validation was performed at this stage to ensure transparency and reproducibility.

Model performance was assessed across multiple metrics to capture goodness-of-fit, predictive accuracy, and generalization capacity. For linear models (SLR and MLR_{EN}), training fit was evaluated using R^2 (SLR) or adjusted $R_{adj}{}^2$–for MLR_{EN}, while machine learning models (ANN, RF, SVM, XGB, and KNN) were assessed using R^2 on training data. Predictive accuracy was quantified using $R_{pred}{}^2$ on validation subsets for ANN, SLR, and MLR_{EN}. Lastly, generalization capacity on external data was tested using $R_{ext}{}^2$ for ANN and MLR_{EN}. Additionally, the root mean square error (RMSE) was calculated for all models to provide an absolute measure of prediction error, and the corrected Akaike information criterion (AIC_c) was applied exclusively for model selection among linear models (SLR and MLR_{EN}) rather than a performance metric. Detailed results are provided in Table S10b.[53]

2.3. Model interpretation

Model interpretation and prediction robustness analyses were conducted on the top-performing models to ensure transparent, reliable and environmentally meaningful predictions. These analyses linked statistical patterns in the data to the underlying physicochemical mechanisms governing PFAS–soil interactions. Predictor importance and contribution were quantified *via* tree-based SHAP that attributes each feature an additive importance value representing its contribution to deviations of individual predictions from the model's mean output. SHAP analysis enables both global interpretations, by ranking predictors according to their mean absolute SHAP values to identify the most influential variables, and local interpretation, by explaining how individual features influence specific predictions while capturing nonlinear and interaction effects.[54]

To evaluate the model sensitivity and quantify uncertainty under realistic environmental variability, Monte Carlo simulations ($N = 5000$) were performed by randomly sampling predictors from their empirical distributions (exponential for SOC, gamma for CEC, and Johnson Sb for SP-3) derived from the PFAS training/validation dataset (Table S12). Simple random sampling was applied without explicitly preserving correlations between predictors, as the primary objective was to quantify the effect of univariate variability or main effects. Specifically, main effects represent the independent contribution of each predictor, assessed by varying predictors individually, while interaction effects were assessed by varying multiple predictors simultaneously to capture interdependencies. Averaging results across iterations provided estimates of both independent (main) and joint (interaction) influences on predicted $\log K_d$.[55,56]

Additionally, model robustness and optimization were conducted through targeted scenario analysis to evaluate model stability under extreme or adverse conditions and to explore optimization strategies. Robustness was tested by simulating boundary-minima predictor values and introducing random noise to predictions, while optimization of predicted $\log K_d$ was performed using a composite desirability function (D; $0 \leq D \leq 1$) representing ideal performance.[57] This approach identified the optimal combination of predictor values (within the observed ranges) maximizing PFAS sorption ($\log K_d$). JMP 18.2 (JMPSD LLC, Cary, NC, USA) was used for data analysis and modeling.

3 Results and discussion

3.1. Data characterization and summary statistics

3.1.1. Characteristics of response variable ($\log K_d$ values for PFAS). The characterization of the training datasets used in this study reveals substantial variability in $\log K_d$ behavior among PFAS subclasses, strongly influenced by molecular structure and functional groups. The distribution of $\log K_d$ values for the carboxylic-PFAS ($N = 327$) was best described by a two-component normal mixture (N2M) provided with the lowest Bayesian information criterion (BIC), indicating the presence of two subpopulations with distinct sorption behaviors, primarily reflecting the differences between short- and long-chain carboxylic PFAS. In contrast, for the PFAS-full dataset ($N = 699$), a three-component normal mixture (N3M) achieved the lowest BIC, consistent with a broader diversity of compounds that exhibit varying sorption mechanisms. This statistical outcome suggests that the PFAS cannot be represented by a single distribution but rather form multiple subpopulations with distinct sorption mechanisms, driven by variations in chain length, degree of fluorination, and functional group chemistry. Across both datasets, $\log K_d$ values spanned a broad range (-1.37 to 3.33 L kg^{-1}) with high variability (CV = 200% and 205%), highlighting the diversity of environmental behaviors among compounds. The right-skewness of the distributions (mean = 0.49 > median = 0.34) suggest that low-sorption compounds are more prevalent, though a smaller subset exhibits markedly higher sorption. The independent external dataset ($N = 628$) displayed similar patterns, and $\log K_d$ values also followed a three-component normal mixture (SD = 0.81) across a similar range (-1.15 to 3.57 L kg^{-1}) but with lower variability (CV = 92%), as shown in Tables S5 and S6, Fig. S1–S3. Collectively, these results reveal

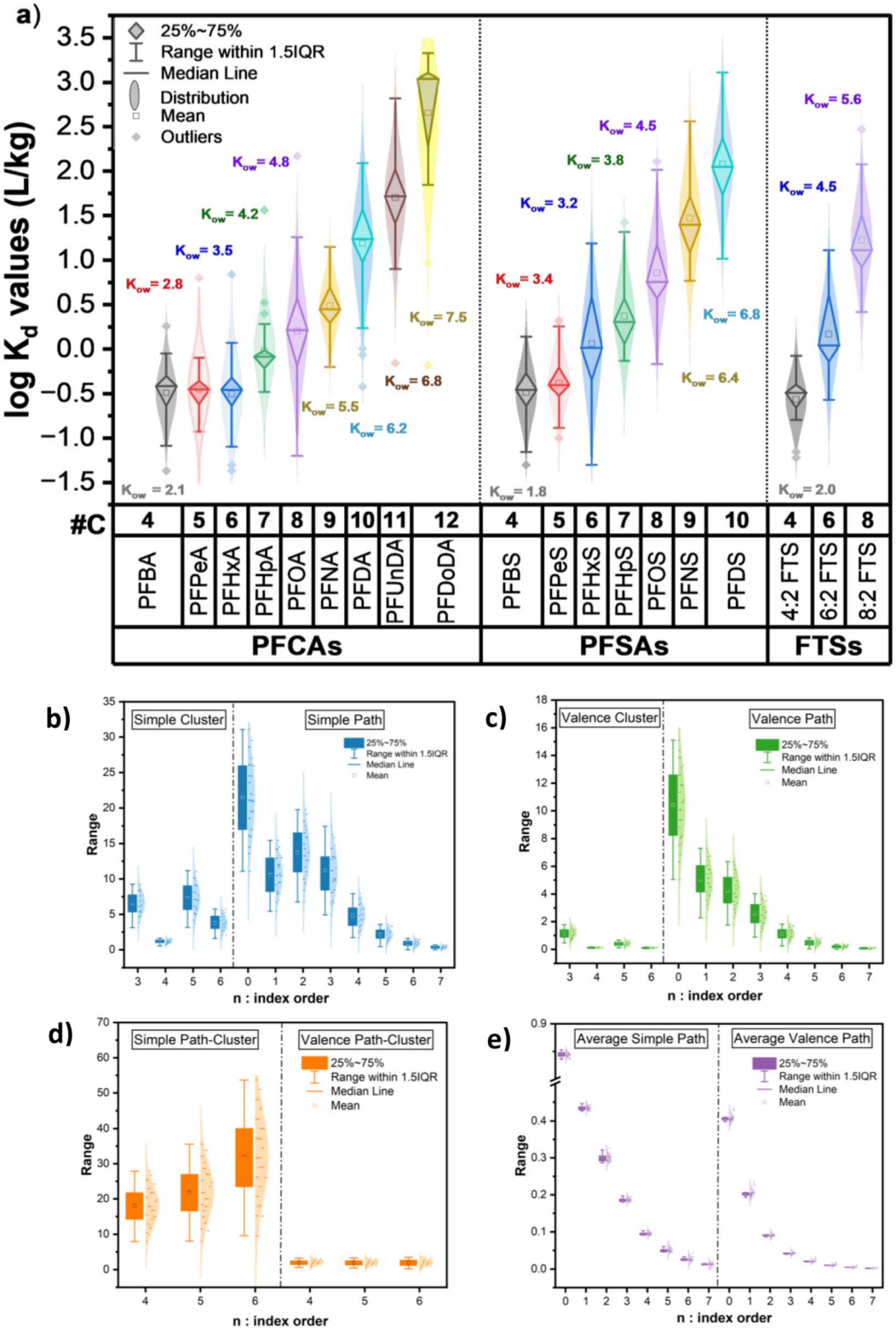

Fig. 4 (a) Distribution and variability plots of experimental log K_d (L kg^{-1}) of PFAS and hydrophobicity (log K_{ow} by the three classes: PFCAs, PFSAs, and FTS) ($N = 699$). (b–e) Box and whisker plots of 46 MCIs grouped by type: simple cluster or path (SC or SP), valence cluster or path (VC or VP), simple and valence path-cluster (SPC and VPC), and average simple path and cluster (ASP and AVP), respectively. Sample size per compound (n): PFBA ($n = 37$), PFPeA ($n = 22$), PFHxA ($n = 35$), PFHpA ($n = 29$), PFOA ($n = 58$), PFNA ($n = 32$), PFDA ($n = 43$), PFUnDA ($n = 40$), PFDoDA ($n = 30$), PFBS ($n = 35$), PFPeS ($n = 32$), PFHxS ($n = 63$), PFHpS ($n = 29$), PFOS ($n = 59$), PFNS ($n = 26$), PFDS ($n = 28$), 4 : 2 FTS ($n = 26$), 6 : 2 FTS ($n = 40$), and 8 : 2 FTS ($n = 34$).

that the PFAS sorption behavior is not uniform but structurally dependent, supporting the need for models that can capture this molecular and environmental complexity.

Fig. 4a illustrates compound-specific log K_d variability within each functional group (*i.e.*, PFCAs, PFSAs, and FTS) and the relationship between C-chain and log K_{ow} (*n*-octanol–water partition coefficient). This figure underscores the pronounced variability not only between subclasses but also among individual PFAS within the same class. For example, PFBA, PFPeA, and PFHxA (C5–C6 carboxylic acids) show narrower distributions with lower median log K_d values, while longer-chain PFCAs like PFOA and PFUnDa (C8–C11) and PFSAs like PFOS and PFDS (C9–C11) exhibit both higher and more dispersed log K_d distributions. Similarly, FTS like 6 : 2 and 8 : 2 FTS display wide distributions. These patterns align with previous observations and reflect that molecular features such as chain length, branching, and functional group chemistry are critical determinants of sorption behavior and the soil property heterogeneity.[9,49,58]

Additionally, a clear positive trend was observed: log K_d increased with the perfluorinated chain length and consequently molecular weight, consistent with the increase in hydrophobicity and stronger sorption affinity to soil. Short-chain PFCAs (C4–C6, MW ≈ 200–400 g mol^{-1}) exhibited narrow log K_d distributions and low median values (median log $K_d \leq 0$ L kg^{-1}), attributed to weak hydrophobic interactions and limited retention by soil.[9] In contrast, long-chain PFCAs (C7–C12, MW ≈ 500–600 g mol^{-1}) and longer-chain PFSAs and FTS demonstrated significantly higher and more variable log K_d values (median log $K_d > 2.0$ L kg^{-1}). However, PFSAs consistently showed higher log K_d values than PFCAs of similar chain lengths, highlighting a stronger affinity of soil for sulfonic acid groups than for carboxylic groups, aligning with previous findings.[35,47] Similarly, FTS generally followed the hydrophobicity trend dictated by the chain length, with 8 : 2 FTS exhibiting high log K_d consistent with its relatively high log K_{ow}. In summary, the shorter chain length resulted in weaker attraction to soil, leading to more consistent log K_d values. The broad data dispersion among the long-chain PFAS reflects greater intermolecular disparity, while variations within individual PFAS classes underscore differences in soil partitioning behaviors influenced by diverse soil properties, as reported in experimental conditions in the related literature.

3.1.2. Characteristics of explanatory variables. The characterization of soil properties in the carboxylic-PFAS ($N = 327$) and PFAS-full ($N = 699$) datasets revealed substantial heterogeneity. Both SOC and CEC followed two- and three-component normal mixture distributions (N2M and N3M), respectively, indicating multiple soil types represented in the data. In the carboxylic-PFAS dataset, SOC ranged from 0.0% to 53.6% (CV = 231%), while in the PFAS-full dataset, it ranged from 0.0% to 7.7% (CV = 120%). The observed zero and near-zero SOC values represent actual measurements reported in the source literature, typically corresponding to mineral-dominated soils with very low organic matter content (*e.g.*, sandy soils).[32] Similarly, CEC ranged from 0.5 to 41.4 cmol kg^{-1} in the carboxylic-PFAS dataset and from 0.0 to 80 cmol kg^{-1} in the PFAS-full dataset, with similar variabilities (CV = 74%) (Table S6), where low or near-zero values also reflect real measurements from soils poor in clay minerals, while higher CEC values correspond to more reactive soils enriched in exchangeable cations (*e.g.*, Ca^{2+}, Mg^{2+}, and Al^{3+}).[32] This property can indirectly influence PFAS sorption by providing more reactive sites,[32,39,41] although its role is typically secondary to the SOC for hydrophobic compounds.[35–37,41] SOC is also a key driver of PFAS sorption due to its ability to facilitate hydrophobic interactions.[30,31] Similarly, the characterization of the independent external dataset, encompassing 35 PFAS, also exhibited pronounced heterogeneity: SOC followed a log-normal distribution with values ranging from 0.03% to 37.6% with high variability (CV = 165%), while CEC adhered to Sinh–Arcsinh (SHASH) distribution with values ranging from 0.10 to 140.0 cmol kg^{-1} with similar variabilities (CV = 111%). Therefore, these analyses confirm that observed low or zero values are genuine field measurements rather than imputed estimates and reflect the natural heterogeneity of soils sampled across diverse environmental conditions.

The soil pH ranged from 3.5 to 8.0 (median pH = ~7), consistent with the conditions spanning acidic to neutral soils. However, the soil pH was not included as a predictor since it was used to account for PFAS speciation (*i.e.*, ionization when the pH exceeds their pK_a) for adjusting the molecular predictors (Text S1). Since PFAS are predominantly anionic, the soil pH influences PFAS speciation.[39,59] This affects both electrostatic interactions and the soil's surface charge because acidic soils reduce electrostatic repulsion and may enhance the sorption of anionic PFAS, while alkaline soils increase repulsion and may reduce sorption.[35]

Fig. 4b–e and S2 show the adjusted predictors (*i.e.*, MCIs and Abraham predictors) that were used in this study. Fig. 4b–e show the distribution of 46 MCIs, which capture structural variability among PFAS. The descriptors were grouped by type—Simple Cluster (SC), Simple Path (SP), Valence Cluster (VC), Valence Path (VP), Simple and Valence Path-Cluster (SPC and VPC), and Average Simple Path and Cluster (ASP and AVP). Substantial variability was observed in lower-order indices (*e.g.*, $n = 1$–3), particularly for simple and valence path descriptors, reflecting their sensitivity to differences in the molecular structure. In contrast, higher-order indices ($n \geq 4$) exhibited narrower ranges, suggesting limited ability to differentiate among compounds. Pearson's correlation matrix analysis of the 46 MCIs and two soil properties revealed the following four main clusters of internally correlated predictors: (1) simple and valence clusters (SC/VC), simple and valence path clusters (SPC/VPC), simple and valence path (SP/VP) ($r > 0.80$ and $p < 0.05$); (2) average simple path (ASP-0 to ASP-5) ($r > 0.60$ and $p < 0.05$); (3) average valence path (AVP) ($r > 0.80$ and $p < 0.05$) and ASP-6/7 ($r = 0.50$ and $p < 0.05$); and (4) soil properties ($r \leq 0.50$ and some p-values > 0.05) (Fig. S4). Cluster 1 was inversely associated with Cluster 2 ($r \leq -0.60$ and $p < 0.05$), while Clusters 3 and 4 showed weaker positive relationships with the others ($r < 0.60$ and $p < 0.05$ and ≤0.9, respectively), indicating the distinct predictor groups. This analysis provided a rationale for MCI-based predictor selection based on the structural heterogeneity of

PFAS in the dataset, prioritizing those combining high variability and moderate complexity for model development.

The Abraham descriptors, calculated for carboxylic PFAS only, also exhibited distinct ranges of variability (Fig. S2). Among these, S' exhibited the widest variability, highlighting substantial differences in the dipolarity/polarizability of PFAS molecules, a property linked to non-specific polar interactions with soil surfaces. E' and V' exhibited moderate ranges, suggesting that excess molar refraction and molecular volume also vary significantly among carboxylic PFAS. Since all PFCAs have the same functional group (*i.e.*, –COOH) in their structure, A', B', and J^- values are similar across the database. These differences in descriptor dispersion are critical for variable selection, as high-variability predictors, such as S', E', and V', are more likely to capture meaningful differences in PFAS sorption behavior.[13]

3.2. Predictors screening and model development

The multi-criteria screening process applied the thresholds for RF-based MCI importance ≥4%, $|r| \geq 0.85$, and $R_{pred}{}^2 \geq 77\%$ in preliminary MLR_{EN} models when paired with the soil properties (Table S7; Fig. S6). This process identified VP-7 and E' for the PFCAs dataset and several MCIs (*i.e.*, VC-3, VC-4, VC-6, VP-1, VP-4, VPC-4, SP-3, SP-4, and SPC-4) as potential top individual predictors for the PFAS dataset. High multicollinearity was detected among the six Abraham descriptors and within the six MCIs (4 SCs, 4 VCs, 3 SPCs, 3 VPCs, 8 VPs, and 8 SPs) compared with ASPs and AVPs (Fig. S4). To mitigate the effects of multicollinearity on model stability and interpretation, we enforced a criterion with VIF ≤ 12. Within-group multicollinearity was resolved by retaining the MCI with the highest $|r|$ to $\log K_d$. Between-group multicollinearity was managed by permitting combinations only if inter-group $|r|$ was below 0.6.

3.2.1. Predictive performance of MCIs for carboxylic PFAS. Before expanding the application of MCIs in a broader PFAS dataset, the predictive potential of MCIs against traditional Abraham descriptors for carboxylic PFAS was tested. The predictor screening process selected a valence path order of seventh MCI (VP-7) and one solvatochromic descriptor (E') as the most relevant predictors. The best-fit MCI model with VP-7 outperformed the best-fit Abraham model with E' across the PFCA training and validation data. Specifically, the MCI model attained a higher $R_{pred}{}^2$ value (84.8%; RMSE = 0.402; and N = 67) than the Abraham model ($R_{pred}{}^2$ = 79.8%; RMSE = 0.469; and N = 76). Both models identified SOC as significant ($P < 0.05$) and CEC as non-significant ($P > 0.05$), with a lower multicollinearity in the MCI model with VP-7 than the Abraham model with E' (VIF ≤ 1.0 *vs.* 1.2–1.4; Table S8). Residual diagnostics confirmed that MLR_{EN} assumptions were adequately met, supporting unbiased coefficient estimates with minimal variance (Fig. S6).

The improved performance of the MCI model is likely attributed to the ability of VP-7 to better capture structural and electronic features, such as molecular size and polarizability, influencing PFAS sorption, than only van der Waals interactions represented by E' in Abraham descriptors.[13] VP-7 is a seventh-order valence path index that quantifies the connectivity of 8 atoms in a linear path and includes electronic information (Table S3).[22] This predictor turned out to be the most important of the 46 MCI tested, indicating that the chain length influenced not only PFAS sorption (*i.e.*, longer C-chains → higher hydrophobicity → higher $\log K_d$; Fig. S7) but also the –COOH group, which has a partial negative charge and a dipole moment that facilitates hydrogen bonding.[22] In other words, we consider that VP-7 spans from the functional group into the tail, encoding how the molecule's polarizability and shape evolve with size. On the other hand, E' captures non-specific interactions,[13,20] which are important in aliphatic compounds such as PFAS, since they do not contain resonating π-electrons but lone-pair electrons.[20,60] Fig. S7b shows that E' values decreased with the increase in PFAS chain length, while $\log K_d$ increased. This inverse trend suggests that the incremental addition of CF_2 units decreases the polarizability of the fluorine tail of the molecule, which decreases its interaction with aqueous phases, facilitating sorption on the solid phase. Besides these mechanistic differences, Abraham descriptors are largely derived from small, neutral organic molecules, potentially limiting their applicability across diverse PFAS chemistries. Therefore, these results highlight the better applicability of MCIs than Abraham descriptors in modeling PFAS sorption for our datasets.

3.2.2. Predictive performance of MCIs across PFAS subclasses. As a performance baseline, the SLR models showed significant correlations between $\log K_d$ and the nine MCIs identified as potential top individual predictors ($p < 0.0001$) (*i.e.*, VP-1, VP-4, VC-3, VC-4, VC-6, VPC-4, SP-3, SP-4, and SPC-4) (Fig. S6). The SLR models with a simple path order 3, SP-3 ($R_{pred}{}^2$ = 71.7%, Model S7), or a valence path order 6, VC-6 ($R_{pred}{}^2$ = 71.5%, Model S5), demonstrated moderate predictive power, highlighting the importance of incorporating multiple MCIs or soil properties (Table S9). The discrepancy between top individual predictors selected by RF in the predictor screening process (VP-4 and VP-1; Table S7) and those selected in SLR models (SP-3 and VC-6; Table S9) highlights the RF's ability to capture non-linear relationships that linear models may overlook. When individual MCIs were combined with the soil properties *via* MLR_{EN}, SP-3 ($R_{pred}{}^2$ = 77.9% for Model M7 in Table S10b) yielded a predictive accuracy improvement of over 6% compared to the best-fit SLR model (Model S7 in Table S9), demonstrating the importance of the soil features. Moreover, MLR_{EN} exhibited reduced overfitting, as indicated by smaller differences between the training and validation RMSE values (0.003–0.044 in Table S10) relative to those in the SLR models (0.007–0.057 in Table S9).

The shift in the most important MCI from VP-7 in the PFCAs dataset to SP-3 in the PFAS dataset is noteworthy. VP-7 performed well with the structurally similar PFCAs, likely capturing subtle electronic differences in the fluorinated chains and head groups. Conversely, SP-3 proved more effective across the diverse 19 PFAS, likely by better differentiating compounds based on the chain length, branching, and overall compactness, irrespective of the specific functional groups. This suggests that higher-order valence indices capture molecular size and electron distribution nuances, crucial for accurate sorption predictions across diverse PFAS.

More complex models incorporating multiple MCIs (Models M10 and A11 in Table S10) were developed by adding two ASP indices and one AVP index to avoid multicollinearity with SP-3, SOC, and CEC. Model M10 evaluated SP-3, ASP-1, ASP-0, and AVP-0 as predictors ($R_{pred}^2 = 83.70\%$; RMSE = 0.413; and VIF ≤ 12), enhancing the predictive performance by >5% over the best-fit single-MCI MLR_{EN} (Model 7b in Table S10). The VIF values (0.8–12 for the MCIs; 0.8–0.9 for the soil properties) remained acceptable (≤12), confirming manageable multicollinearity[51] (Table S10).

Equivalent six-predictor machine learning models were evaluated, resulting in ANN outperforming its counterpart RF ($R_{pred}^2 = 82.42\%$; RMSE 0.428; and $N = 175$), support vector machine ($R_{pred}^2 = 82.14\%$; RMSE 0.432; and $N = 164$), extreme gradient boosting ($R_{pred}^2 = 78.37\%$; RMSE 0.475; and $N = 175$), and K nearest neighbors ($R_{pred}^2 = 78.35\%$; RMSE 0.475; and $N = 175$) models. The ANN model using SP-3, ASP-1, ASP-0, AVP-0, SOC, and CEC as predictors (Model A11 in Table S10; a single hidden layer; three neurons; the hyperbolic tangent—Tan H—activation function; 20 boosting iterations; Fig. S5) yielded a higher predictive accuracy ($R_{pred}^2 = 84.9\%$; RMSE = 0.397; and $N = 164$) than MLR_{EN} ($R_{pred}^2 = 83.7\%$; RMSE 0.413; and $N = 164$). This finding suggests that the ANN model captured non-linearities and predictor interactions that the linear models missed. However, this increased complexity might not translate into better generalization. Table 1 summarizes the performance metrics of the best predictive models; the italic-row (Model M7) indeed represents our selected, most generalized model, chosen based on its optimal balance between high predictive accuracies (R_{pred}^2) on validation datasets.

To contextualize our results, the modeling framework and predictive performance of our model were compared qualitatively with a few recent PFAS sorption modeling studies.[49,61,62] As summarized in Table S14, our study stands out by using MCIs as PFAS descriptors. Unlike conventional physicochemical or geometric descriptors, such as molecular weight, hydrophobicity, solubility, or molecular size, MCIs represent atomic connectivity and branching patterns within the molecular graph. This allows them to encode structural information without relying on 3D conformations or experimentally derived physicochemical parameters, making them computationally efficient and broadly applicable across diverse PFAS structures. The predictive accuracy obtained in our study ($R^2 = 77.9$–84.9% and RMSE = 0.40–0.50) achieved comparable R^2 values (72–93%) and similar RMSE ranges (0.36–0.86), underscoring the robustness and transferability of the MCI-based modeling framework. However, direct numerical comparison across studies is not strictly appropriate, as the reported models relied on different descriptor types (*e.g.*, molecular weight, $\log K_{ow}$, and charge density) and distinct machine learning frameworks (*e.g.*, RF and LGBM). Furthermore, the underlying soil datasets differ substantially among studies (Table S15). Our dataset encompasses soils with relatively low mean SOC (1.13%) and moderate CEC (16.3 cmol kg^{-1}) compared with prior works (SOC = 2–4.7%; CEC = 16–19 cmol kg^{-1}), while maintaining a comparable pH range (6–7). These variations in soil composition and experimental conditions make strict side-by-side comparison difficult but nonetheless highlight the practical advantage of MCIs as versatile predictors to evaluate the PFAS sorption behavior ($\log K_d$).

3.3. Model evaluation

3.3.1. Predictor's importance and interaction effects. The feature importance of predictors derived from the models was quantified by SHAP values that explain and compare how individual predictors affect model outputs, capturing both the direction and magnitude of influence across Models M7, M10, and A11 (Fig. 5). The SHAP summary plot for Model M7 (Fig. 5a) highlights SP-3 as the most influential predictor, with its low and high values spanning SHAP contributions from approximately −1.5 to 1.5. High SP-3 values (red) corresponded to increased predicted $\log K_d$ (positive SHAP values), indicating that more complex structures may enhance PFAS sorption. Fig. S9 illustrates a perfect correlation between SP-3 and molecular weight across all PFAS subclasses ($R^2 = 1.00$), confirming that SP-3 captured molecular size and chain structure, two primary drivers of PFAS sorption. This correlation also indicates that SP-3 reflects PFAS hydrophobicity. Therefore, SP-3 provides a unified metric that bridges both structural and physicochemical factors affecting sorption, explaining its superior performance in the PFAS-full dataset.

Table 1 Model performance metrics of the best predictive performance models with the most important predictors[a]

	Model type	Model No.	Predictor importance (highest → lowest)	T		V	
				R_{adj}^2 (%)	RMSE	R_{pred}^2 (%)	RMSE
PFCAs	Abraham descriptors	—	E, SOC, and CEC	80.2	0.51	79.8	0.47
	MCIs descriptors	—	VP-7, SOC, and CEC	82.9	0.47	84.8	0.40
3 PFAS subclasses	MCIs descriptors	*M7*	*SP-3, SOC, and CEC*	*77.1*	*0.478*	*77.9*	*0.481*
		M10	SP-3, ASP-1, ASP-0, AVP-0, SOC, and CEC	84.4	0.394	83.7	0.413
		A11	SP-3, ASP-1, AVP-0, ASP-0, SOC, and CEC	86.3*	0.369	84.9	0.397

[a] V: validation and T: training. *R^2: goodness-of-fit on training data in A11. R_{pred}^2: goodness-of-fit on validation data and R_{adj}^2: goodness-of-fit on training data. RMSE: root mean square error.

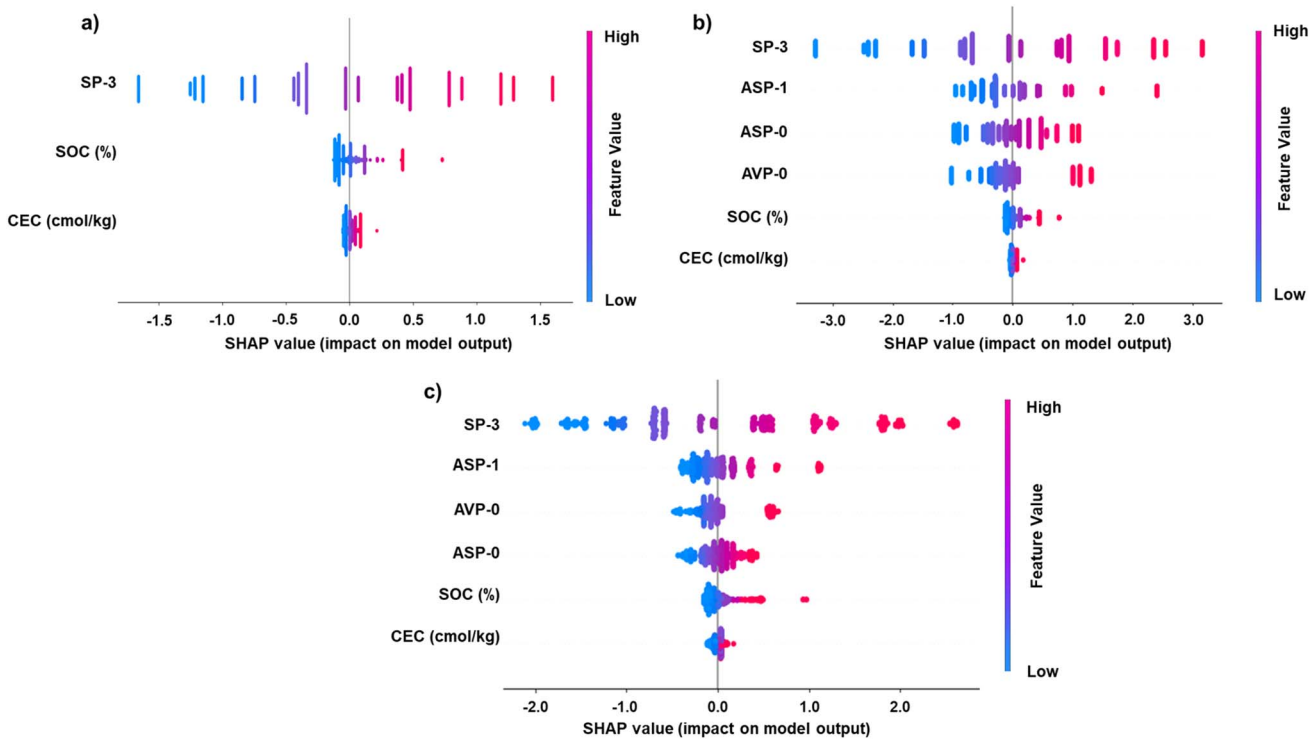

Fig. 5 SHAP summary plots showing the influence of individual features on log K_d predictions for (a) M7, (b) M10, and (c) A11 models. Features are ranked from top to bottom by their relative importance in determining the model output. Each dot represents a single sample, where the color gradient from blue (●, low) to red (●, high) indicates the feature value. The horizontal position represents the SHAP value (in log K_d units), which quantifies each feature's contribution to deviations from the model's mean prediction (baseline). Negative SHAP values indicate a lowering effect on the predicted log K_d, while positive SHAP values indicate an increasing effect. Axes are not standardized across panels and variations in SHAP ranges reflect differences in the magnitude and distribution of feature contributions across models.

Similarly, SP-3 remained the predominant positive contributor in the SHAP plot of Models M10 and A11, as illustrated in Fig. 5b and c, although both had broader SHAP values than Model M7, reflecting their increased complexity. While SP-3 remained the dominant positive contributor, the inclusion of ASP-1, ASP-0, and AVP-0 introduced greater variance in predictor effects. These three low-variance variables disproportionately contributed to SHAP space—narrow but impactful SHAP bands and non-monotonic relationships (Fig. 5b). In other words, this model was sensitive to small differences in these variables, suggesting potential overfitting to dataset-specific noise.

In Model A11, as seen in the SHAP plot (Fig. 5c), the non-linear nature of the ANN model was manifested as more complex and asymmetric distributions. SP-3 again exerted the largest influence, with its high values (red) correlating with higher log K_d. However, ASP-1, ASP-0, and AVP-0 still contributed non-trivially despite their minimal variability, indicating that the ANN captured interactions between the low-variance MCIs and dynamic soil features. However, the wide SHAP fluctuations for ASP-1 and AVP-0 highlighted the model's high sensitivity to low-signal inputs and risk of model instability. In contrast, SOC and CEC appeared to exert less overall impact on the model than the molecular descriptors. While low SOC values were associated with negative SHAP values (reduced sorption potential), their limited spread suggests a secondary influence relative to molecular features, aligning with previous findings.[49,58] Similarly, CEC showed a relatively narrow effect range, with low CEC values contributing negatively to predictions. Overall, the model indicates that PFAS sorption was driven more by intrinsic molecular features, particularly connectivity-based descriptors like SP-3, than by soil characteristics in the current dataset.[49,58] These results underscore the utility of MCIs in capturing the structural determinants of PFAS behavior in the environment.

Monte Carlo simulations ($N = 5000$) were leveraged to further decompose the total importance of each predictor into main (predictor independency) and interaction (predictor dependency) effects on log K_d predictions within the PFAS data context. Across all three models, the general order of predictor importance was MCIs > SOC > CEC. Both the non-linear (ANN) and linear (MLR_{EN}) approaches converged on the primacy of SP-3, followed by ASP-1, but diverged on the relative importance of ASP-0 *versus* AVP-0 (ASP-0 > AVP-0 for Model M10; AVP-0 > ASP-0 for Model A11). The main effects accounted for 94–97% of the predictive power of Models M7, M10, and A11, whereas interaction effects contributed only 3–6%. Among all predictors, SP-3 had the largest share of the main effects raging from 68 to 95%, with minimal interaction contributions (1–2%). Other MCIs exhibited substantially smaller main effects, including ASP-1

Table 2 Evaluation of the external predictive performance of three models with the best performance in their training and validation datasets[a]

		18 PFAS subsets				35 PFAS subsets			
		Soil-only		Soil and sediments		Soil-only		Soil and sediments	
Model no.	MCI predictor and soil properties	R_{ext}^2 (%)	RMSE	R_{ext}^2 (%)	RMSE	R_{ext}^2 (%)	RMSE	R_{ext}^2 (%)	RMSE
M7	*SP-3, SOC, and CEC*	*49.6*	*0.55*	*53.1*	*0.56*	*45.6*	*0.82*	*52.4*	*0.62*
M10	SP-3, ASP-1, ASP-0, AVP-0, SOC, and CEC	47.5	0.61	51.4	0.62	20.2	1.22	19.2	1.01
A11	SP-3, ASP-1, AVP-0, ASP-0, SOC, and CEC	4.87	3.62	7.33	3.61	5.37	6.86	29.4	1.15

[a] 18 PFAS subsets: same families as those in the training and validation datasets (PFCAs, PFSAs, and FTS), except for 4 : 2 FTS. 35 PFAS subsets: representing a broader and more diverse set of chemical subclasses and emerging compounds. R_{ext}^2: goodness-of-fit on external data. RMSE: root mean square error.

(6–13%), ASP-0 (2–7%)), and AVP-0 (4–6%), each with 1% interaction contributions. Soil properties contributed modestly to predictive power, SOC accounting for 1–2% of the main effect and 0.4–1.1% of the interaction effect, while CEC explained only 0.02–0.2% of the main effect and 0.02–0.1% of the interaction effect. Overall, interaction effects were minimal across all predictors, indicating that model predictions were primarily driven by additive main effects rather than predictor interactions. The ANN's ability to capture these subtle interactions likely contributed to its higher validation R_{pred}^2 than that of the inherently linear MLR_{EN}. However, the generalization capacity of the models remains to be tested to determine if these captured patterns are mechanistically plausible and generalized. Across all the MLR_{EN} models, the rates of change in $\log K_d$ with a one-unit increase ranged from 0.16 ± 0.004 (Model M9; SPC-4) to 123.7 ± 7.5 (Model M10; ASP-1) for the MCIs ($P < 0.0001$), from 0.098 ± 0.016 (Model M1) to 0.12 ± 0.012 (Model M10) for SOC ($P < 0.0001$), and from 0.002 ± 0.001 (Models M1, M2, and M10) to 0.004 ± 0.001 (Models M9) for CEC ($P < 0.05$ in only four Models M3, M5, M6, M9) (Table S10a). The large positive slopes indicated high sensitivity of $\log K_d$ to small changes in ASP-1 (123.7), ASP-0 (122.00), AVP-0 (51.5), VC-4 (23.0), and VC-6 (22.0) (Table S10a).

3.3.2. External predictive performance. To assess the generalizability of the models, external validation was conducted using two datasets: (1) soil-only datasets and (2) a combined soil & sediments dataset (Fig. 3). These were tested against two PFAS subsets: one with 18 compounds (similar to the 19 PFAS-full training/validation dataset) and another with 35 PFAS encompassing broader chemical families. This evaluation is essential, as real-world utility depends not only on performance with familiar data but also on the model's ability to predict outcomes in distributionally distinct datasets. Compared to the holdout validation set of 19 PFAS, the external datasets included a wider range of predictors, more chemically diverse PFAS, and more variable distributions (Table S6), offering a rigorous test of model performance.

Table 2 summarizes the external predictive performance of models M7, M10, and A11 tested for the soil-only datasets and soil & sediments datasets. Across all cases, the model's performance increased when the combined soil & sediments datasets were tested. Specifically, Model M7 (italic-row) consistently achieves the highest R^2 values and lowest RMSEs, particularly for the soil & sediments datasets (18 PFAS subset: $R^2 = 53.10\%$, RMSE = 0.56; 35 and PFAS subset: $R^2 = 52.40\%$, RMSE = 0.62), suggesting its robustness and applicability to more chemically diverse and environmentally complex datasets in comparison with Models M10 and A11. These findings emphasize that validation serves as a necessary condition for model viability, while external validation represents sufficient condition for real-world predictability: both are indispensable for robust forecasting. Moreover, the inclusion of sediment data appears to enhance the model performance by increasing the diversity and representativeness of predictor distributions, thus better capturing the complexity of PFAS sorption behavior. This is particularly important given the limited data available on the adsorption and desorption behaviors of long-chain and emerging PFAS in sediments.[63–66]

The underlying cause of this performance divergence was three-fold, stemming from the interplay between the data characteristics and model complexity. First, all the six predictors and $\log K_d$ showed significant distributional shifts between the 19 PFAS (training/validation) and external validation data (Table S6). Second, SOC and CEC demonstrated greater variability in the external data (35 PFAS soil & sediments) than in the 19 PFAS data, given the values of SD (CEC: 26.39 *vs.* 11.98; SOC: 5.25 *vs.* 1.36), CV (CEC: 111 *vs.* 74; SOC: 165 *vs.* 121), range (CEC: 140.00 *vs.* 80.00; SOC: 37.60 *vs.* 7.70), and IQR (CEC: 18.00 *vs.* 14.50; SOC: 2.80 *vs.* 1.15) (Table S6). All the models trained on the narrower PFAS dataset ranges struggled to accurately generalize to these broader external conditions. Third, MCIs such as ASP-0, ASP-1, and AVP-0 displayed extremely low variance across all the datasets (CV = 1–4%), limiting their discriminatory power. The linear or non-linear exploitation of their non-generalizable fluctuations by Models M10 and A11, respectively, likely inflated their predictive accuracy but reduced their generalization capacity. In contrast, Model M7 avoided these pitfalls by focusing on SOC, CEC, and SP-3 with the substantial variance and stable influence—as also shown by the stable predictive accuracy of three-predictor (similarly

MCI predictive equation:

$$\log K_d = 0.26 \times SP_3 + 0.11 \times SOC + 0.003 \times CEC.$$

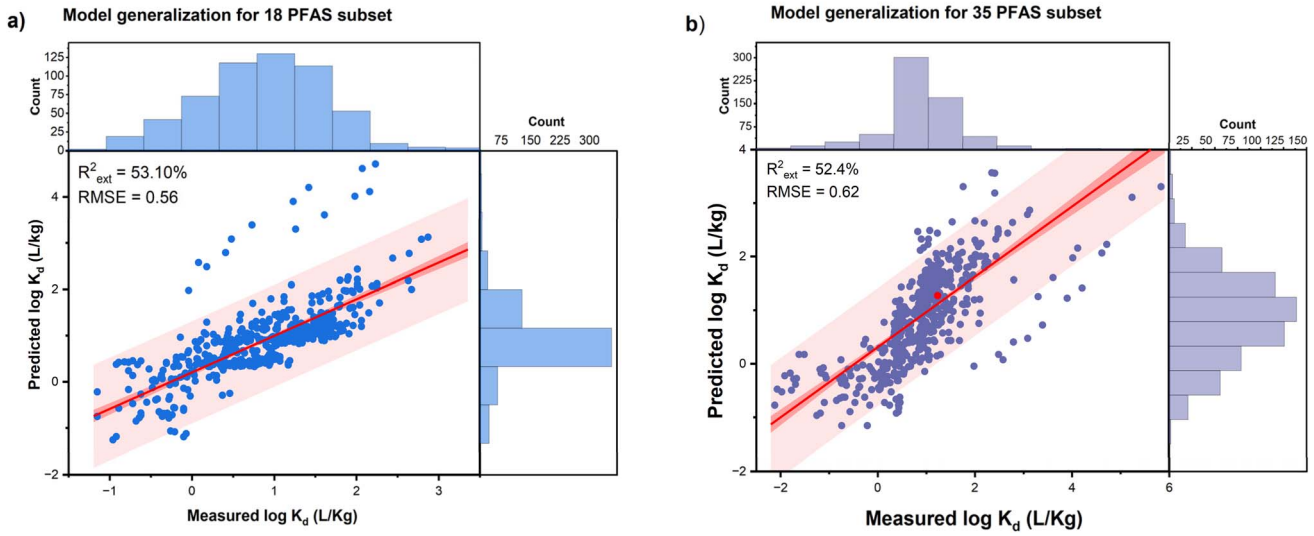

Fig. 6 Scattered plots and histograms displaying the distribution of measured *vs.* predicted log K_d (L kg^{-1}) values of model M7 tested with the external soil & sediments dataset for (a) 18 PFAS subsets and (b) 35 PFAS subsets. The dark red regression line illustrates the actual trend between the predicted and measured values. The light and darker red shaded bands represent the 95% confidence and prediction intervals, respectively, surrounding the regression line.

parsimonious) Models M3–M9 (R_{pred}^2 = 76.78–77.91; RMSE = 0.481–0.493 in Tables 1 and S10b). Thus, the predictive gains (increased complexity) from adding the multiple low-variance MCIs to Models M10 and A11 (despite meeting VIF criteria) did not translate into better generalization, thus highlighting the trade-off between model complexity and real-world reliability.

It is important to mention that our study and a recent publication[61] used independent external datasets for evaluating the external generalization of the models, resulting in R_{ext}^2 values of 52.4% and 71.7%, respectively. In our case, prediction bias within the external dataset was primarily associated with samples exhibiting higher mean and median log K_d values, indicative of stronger PFAS adsorption. This pattern likely reflects greater variability and higher mean SOC, suggesting soils richer in organic matter compared to the training and validation datasets. Similarly, CEC values in the external dataset displayed higher averages and extreme ranges, pointing to more diverse soil mineralogy and an abundance of charge sites (Table S6). Overall, the external dataset was more heterogeneous and encompassed more complex soil conditions, with elevated SOC and CEC correlating positively with higher log K_d. These findings indicate that our model performs more accurately for soils containing low-to medium-sorbing PFAS, while prediction becomes more challenging under conditions of high organic carbon and cation exchange capacity.

Fig. 6 presents the regression equation of the highest external predictive performance across both soil-only and soil & sediments datasets (Table 2). The coefficients illustrate the combined roles of molecular connectivity index (SP-3), organic matter content (SOC), and cation exchange capacity (CEC) in controlling PFAS partitioning. SP-3, the dominant predictor, captures the effects of molecular connectivity, chain length, and hydrophobic surface area, which are structural attributes known to enhance sorption. SOC exerts a moderate positive effect, consistent with the established role of organic matter in promoting hydrophobic partitioning of PFAS.[32] Although the CEC coefficient is comparatively small, it remains statistically significant and improves model stability and generalization by representing soil cation-exchange and ion-bridging mechanisms.[32] Collectively, these predictors describe complementary molecular and environmental controls governing the PFAS sorption behavior.

3.3.3. Scenario analyses. Scenario analyses explored: (i) the optimization of maximal log K_d—corresponding to minimal PFAS mobility in environmental matrices such as soils and sediments—across all the 11 models *via* a composite desirability function (*D*), (ii) model uncertainty, and (iii) model robustness or stability under boundary minima *via* Monte Carlo simulations. Maximal PFAS sorption yielded a consistent positive correlation of log K_d with both the soil and MCI predictors (Table S11). However, unlike the three-predictor models (*e.g.*, Model M7), the six-predictor models (*e.g.*, Models M10 and S11) produced extrapolated log K_d predictions beyond their observed ranges, verifying their poor generalization. Although the single-MCI MLR_{EN} models yielded slightly lower maximum desirability scores (0.83–0.89) than the multiple-MCIs models (0.99), their superior generalization makes them more suitable for realistic

environmental applications aimed at limiting PFAS transport. Notably, the soil conditions that yielded the highest predicted PFAS retention (SOC = 7.7% and CEC = 80 cmol kg^{-1}) exceed typical background soil ranges (SOC < 5% and CEC < 30 cmol kg^{-1}),[67] highlighting the gap between optimal model predictions and practical field conditions.

To propagate predictor uncertainty through Model M7, Monte Carlo simulations ($N = 5000$) resampling SOC, CEC, and SP-3 from exponential ($\sigma = 1.119$), gamma (shape = 1.66; scale = 9.51), and Johnson Sb ($\theta = 1.20$; $\delta = 1.50$; scale = 20.3) distributions, respectively, were performed (Table S12). These best-fit distributions were selected based on their PFAS (training/validation) dataset. Both Model M7 predictions ($N = 638$) and simulations ($N = 5000$) exhibited a three-component normal mixture distribution, reflecting the inherent complexity of PFAS sorption across the diverse environmental regimes. The simulations produced slightly higher mean (0.56 *vs.* 0.52) and median (0.56 *vs.* 0.50) $\log K_d$ values, with reduced dispersion (SD: 0.81 *vs.* 0.88; IQR: 1.13 *vs.* 1.23), than the deterministic predictions of Model M7 (Table S12).

The simulations also exhibited an expanded prediction range (−1.65, 2.91) relative to that of Model M7 (−1.28, 2.56), revealing the underestimation of PFAS mobility/retention extremes by the model. These extremes are far more critical for risk management than central estimates. Therefore, there is a need for closer scrutiny of the three-component normal mixture distribution to account for high-impact tail behaviors. The low-sorption component (indicative of high PFAS mobility) on average shifted from −0.85 (95% CI: −0.91 and −0.79) to −0.28 (95% CI: −0.31 and −0.26), with variability doubling (σ: 0.26 → 0.50) and proportion tripling (π: 0.11 → 0.31). This suggests that PFAS mobility risks are higher under real-world conditions than predicted by the central tendencies of Model M7.[68] Meanwhile, the mid-sorption component (indicative of moderate PFAS sorption) moved from 0.54 (95% CI: 0.48, 0.60) to 0.69 (95% CI: 0.67, 0.70), with tighter dispersion (σ: 0.70 → 0.46) and lower dominance (π: 0.79 → 0.51). The high-sorption component (indicative of PFAS retention in soil) decreased from 1.80 (95% CI: 1.75, 1.85) to 1.63 (95% CI: 1.60, 1.66), with increased spread (σ: 0.21 → 0.42) and higher probability (π: 0.11 → 0.18). In particular, the combined probability of extreme (low sorption + high sorption) outcomes nearly doubled from 21.5% to 49.2% in the simulation results (Table S12). In other words, the deterministic model (Model M7) underestimated extreme risks (PFAS mobility/retention) by approximately 184% (low-sorption) and 71% (high-sorption). These shifts underscore how deterministic models may significantly underestimate both the frequency and variability of extreme PFAS behaviors. This is a critical consideration for risk management and remediation prioritization, as underestimating low-sorption (high PFAS mobility) scenarios can cause inadequate containment strategies and greater environmental exposure risk, while overestimating high-sorption may lead to insufficient remediation due to inaccurate estimations. The incorporation of predictor uncertainty enables more balanced and informed decision-making, particularly when planning safeguards for high-impact, low-frequency outcomes.

Finally, boundary-minima robustness (stability) evaluates how well Model M7 performs under extreme, low-sorption conditions (representing a worst-case scenario for PFAS mobility) by fixing the predictors at their minimum observed values (SOC: 0.1%; CEC: 0.5 cmol kg^{-1}; and SP-3: 4.9) and adding Gaussian noise (SD = 0.88) to $\log K_d$ across 5000 simulations. The simulated mean (−1.30; 95% CI: −1.42 and −1.19) closely matched the model's predicted minimum (−1.30; 95% CI: −1.33 and −1.28), confirming unbiased predictions under extreme low-sorption conditions. The simulated SD (0.89) closely matched the noise magnitude (SD = 0.88), demonstrating linear and predictable error propagation and model stability at this operational boundary. The narrow 95% CI (0.05) further underscored the model's precision at this operating point. In other words, the model intercept and slopes gave a reliable estimate at this boundary, even under the noisy $\log K_d$ measurement.

In summary, these scenario analyses reaffirm Model M7's stability for the central and boundary predictions while emphasizing the need for probabilistic outputs in actionable risk assessments, given its deterministic nature. To capture the full range of plausible outcomes, future studies should explore dynamic modeling to incorporate spatiotemporal variability in soil and PFAS properties. Embedding such dynamics into decision-support tools can help balance remediation costs with ecological and public health risks.

4 Conclusions

This study advances PFAS sorption modeling by integrating molecular connectivity indices (MCIs) with soil properties to predict $\log K_d$ across chemically diverse PFAS. Compared with prior QSAR and machine learning models, the proposed framework achieves comparable generalization while using a more parsimonious and interpretable predictor set. This parsimony enhances model transparency and practical usability in data-limited scenarios where detailed molecular descriptors or physicochemical measurements are unavailable. In particular, the most influential descriptor, SP-3, effectively captured dominant molecular determinants of sorption, accounting for 68–95% of main-effect variance through structural connectivity and hydrophobicity. The inclusion of soil organic carbon (SOC) and cation exchange capacity (CEC) further improved external stability and mechanistic realism, representing hydrophobic partitioning and cation-bridging pathways, respectively. Although CEC exhibited a smaller numerical coefficient, it remained statistically significant and improved the generalization performance, reinforcing its mechanistic relevance for anionic PFAS.

While the three-predictor MLR_{EN} model (Model M7) exhibited moderate external generalization ($R_{ext}^2 = 52.4\%$) relative to internal validation ($R_{pred}^2 = 77.9\%$), its balanced combination of interpretability, simplicity, and robustness makes it a practical screening-level framework for rapid PFAS sorption assessment. The ANN models achieved higher accuracy during training/validation by capturing non-linear interactions but displayed reduced external reliability, highlighting the trade-off between model complexity and generalizability. Model predictions should therefore be interpreted within the context of

uncertainty and chemical domain coverage, particularly for underrepresented PFAS subclasses.

Future research should focus on improving model generalizability and applicability by expanding the PFAS and soil datasets, incorporating additional physicochemical descriptors, soil composition variables, and spatiotemporal dynamics to better reflect the complexity of PFAS behavior and further enhance the model's generalizability, robustness, and potential applications in environmental risk assessment.

Author contributions

Alulema-Pullupaxi P.: formal analysis, data collection, writing—original draft and editing; Evrendilek F.: methodology, writing—original draft and editing; Hatinoglu D.: formal analysis, data collection, review and editing; Moavenzadeh Ghaznavi S.: data collection, review and editing; Mensah K.: data collection, review and editing; Choudhary M.: data collection, review and editing; Ortiz S.: data collection, review and editing; and Apul O.: conceptualization, supervision, writing—original draft and editing.

Conflicts of interest

The authors declare no competing interests.

Abbreviations

AICc	Corrected Akaike information criterion
ANN	Artificial neural network
C#	Number of carbon atoms in the molecule (*e.g.*, C4 or C8)
CEC	Cation exchange capacity
FTS	Fluorotelomer sulfonates
KNN	K-nearest neighbor
K_{oa}	Octanol–air partitioning coefficient
K_{ow}	Octanol–water partitioning coefficient
Log K_d	Soil–water partitioning coefficient
LSER	Linear solvation energy relationship
MCIs	Molecular connectivity indices
ML	Machine learning
N	Sample size
PFAS	Per- and poly-fluoroalkyl substances
PFCAs	Perfluorocarboxylic acids
PFSAs	Perfluorosulfonic acids
QSPR	Quantitative structure–property relationship
QSAR	Quantitative structure–activity relationship
R_{adj}^2	Adjusted coefficient of determination on training dataset
R_{pred}^2	Predicted coefficient of determination on validation dataset
R_{ext}^2	Predicted coefficient of determination on independent external dataset
RF	Random forest
SHAP	Shapley additive explanations
SMILES	Simplified molecular-input line-entry system
SOC	Soil organic carbon
SP-3	Simple-path order 3
SVM	Support vector machine
VIF	Variance inflation factor
VP-7	Valence-path order 7
XGB	Extreme gradient boosting

Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information: 37 pages, 15 tables, and 9 figures. See DOI: https://doi.org/10.1039/d5em00532a.

Acknowledgements

This work was supported by the NSF (Award 2449798 and 2219832) and the USDA PFAS NACA Grant (58-8030-4-015) at the University of Maine. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

1 M. Hannan, F. Evrendilek, D. Leclair, M. Choudhary, K. Mensah, C. Aeppli, A. K. Venkatesan and O. G. Apul, Aftermath of a Major Firefighting Foam Spill in Brunswick, Maine: Spatiotemporal Dynamics of per- and Polyfluoroalkyl Substances in the Downstream Surface Waters, *J. Hazard. Mater. Lett.*, 2025, **6**, 100150, DOI: 10.1016/j.hazl.2025.100150.

2 C. Abou-Khalil, D. Sarkar, P. Braykaa and M. C. Boufadel, Mobilization of Per- and Polyfluoroalkyl Substances (PFAS) in Soils: A Review, *Curr. Pollut. Rep.*, 2022, **8**(4), 422–444, DOI: 10.1007/s40726-022-00241-8.

3 S. Moavenzadeh Ghaznavi, C. Zimmerman, M. E. Shea, J. D. MacRae, J. M. Peckenham, C. L. Noblet, O. G. Apul and A. D. Kopec, Management of Per- and Polyfluoroalkyl Substances (PFAS)-Laden Wastewater Sludge in Maine: Perspectives on a Wicked Problem, *Biointerphases*, 2023, **18**(4), DOI: 10.1116/6.0002796.

4 M. Mudlaff, A. Sosnowska, L. Gorb, N. Bulawska, K. Jagiello and T. Puzyn, Environmental Impact of PFAS: Filling Data Gaps Using Theoretical Quantum Chemistry and QSPR Modeling, *Environ. Int.*, 2024, **185**, 108568, DOI: 10.1016/j.envint.2024.108568.

5 J. Smith, M. L. Brusseau and B. Guo, An Integrated Analytical Modeling Framework for Determining Site-Specific Soil Screening Levels for PFAS, *Water Res.*, 2024, **252**, 121236, DOI: 10.1016/j.watres.2024.121236.

6 M. W. Sima and P. R. Jaffé, A Critical Review of Modeling Poly- and Perfluoroalkyl Substances (PFAS) in the Soil-Water Environment, *Sci. Total Environ.*, 2021, **757**, 143793, DOI: 10.1016/j.scitotenv.2020.143793.

7 M. Hubert, H. P. H. Arp, M. C. Hansen, G. Castro, T. Meyn, A. G. Asimakopoulos and S. E. Hale, Influence of Grain Size, Organic Carbon and Organic Matter Residue Content on the Sorption of per- and Polyfluoroalkyl Substances in Aqueous Film Forming Foam Contaminated Soils -

Implications for Remediation Using Soil Washing, *Sci. Total Environ.*, 2023, **875**, 162668, DOI: 10.1016/j.scitotenv.2023.162668.

8 G. Feng, B. Zhou, R. Yuan, S. Luo, N. Gai and H. Chen, Influence of Soil Composition and Environmental Factors on the Adsorption of Per- and Polyfluoroalkyl Substances: A Review, *Sci. Total Environ.*, 2024, **925**, 171785, DOI: 10.1016/j.scitotenv.2024.171785.

9 T. M. H. Nguyen, J. Brä, K. Thompson, J. Thompson, S. Kabiri, D. A. Navarro, R. S. Kookana, C. Grimison, C. M. Barnes, C. P. Higgins, M. J. Mclaughlin and J. F. Mueller, Influences of Chemical Properties, Soil Properties, and Solution PH on Soil−Water Partitioning Coefficients of Per-and Polyfluoroalkyl Substances (PFASs), *Environ. Sci. Technol.*, 2020, **54**, 15883–15892, DOI: 10.1021/acs.est.0c05705.

10 J. Jeon, K. Kannan, B. J. Lim, K. G. An and S. D. Kim, Effects of Salinity and Organic Matter on the Partitioning of Perfluoroalkyl Acid (PFAs) to Clay Particles, *J. Environ. Monit.*, 2011, **13**(6), 1803, DOI: 10.1039/c0em00791a.

11 J. L. Rayner, D. Slee, S. Falvey, R. Kookana, E. Bekele, G. Stevenson, A. Lee and G. B. Davis, Laboratory Batch Representation of PFAS Leaching from Aged Field Soils: Intercomparison across New and Standard Approaches, *Sci. Total Environ.*, 2022, **838**, 156562, DOI: 10.1016/j.scitotenv.2022.156562.

12 R. Jin, Y. Liang and Z. Shi, Machine Learning Prediction of DOC–Water Partitioning Coefficients for Organic Pollutants from Diverse DOM Origins, *Environ. Sci. Process. Impacts*, 2025, DOI: 10.1039/D5EM00029G.

13 M. D. Hatinoglu, F. Perreault and O. G. Apul, Modified Linear Solvation Energy Relationships for Adsorption of Perfluorocarboxylic Acids by Polystyrene Microplastics, *Sci. Total Environ.*, 2023, **860**, 160524, DOI: 10.1016/j.scitotenv.2022.160524.

14 Z. Wang, M. MacLeod, I. T. Cousins, M. Scheringer and K. Hungerbühler, Using COSMOtherm to Predict Physicochemical Properties of Poly- and Perfluorinated Alkyl Substances (PFASs), *Environ. Chem.*, 2011, **8**(4), 389, DOI: 10.1071/EN10143.

15 M. L. Brusseau and S. Van Glubt, The Influence of Molecular Structure on PFAS Adsorption at Air-Water Interfaces in Electrolyte Solutions, *Chemosphere*, 2021, **281**, 130829, DOI: 10.1016/j.chemosphere.2021.130829.

16 M. L. Brusseau, The Influence of Molecular Structure on the Adsorption of PFAS to Fluid-Fluid Interfaces: Using QSPR to Predict Interfacial Adsorption Coefficients, *Water Res.*, 2019, **152**, 148–158, DOI: 10.1016/j.watres.2018.12.057.

17 T. M. Nolte and A. M. J. Ragas, A Review of Quantitative Structure–Property Relationships for the Fate of Ionizable Organic Chemicals in Water Matrices and Identification of Knowledge Gaps, *Environ. Sci. Process. Impacts*, 2017, **19**(3), 221–246, DOI: 10.1039/C7EM00034K.

18 C. I. Cappelli, E. Benfenati and J. Cester, Evaluation of QSAR Models for Predicting the Partition Coefficient (LogP) of Chemicals under the REACH Regulation, *Environ. Res.*, 2015, **143**, 26–32, DOI: 10.1016/j.envres.2015.09.025.

19 J. W. Washington, T. M. Jenkins, K. Rankin and J. E. Naile, Decades-Scale Degradation of Commercial, Side-Chain, Fluorotelomer-Based Polymers in Soils and Water, *Environ. Sci. Technol.*, 2015, **49**(2), 915–923, DOI: 10.1021/es504347u.

20 O. G. Apul, F. Perreault, G. Ersan and T. Karanfil, Linear Solvation Energy Relationship Development for Adsorption of Synthetic Organic Compounds by Carbon Nanomaterials: An Overview of the Last Decade, *Environ. Sci.*, 2020, **6**(11), 2949–2957, DOI: 10.1039/D0EW00644K.

21 L. H. Hall, and L. B. Kier, The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling, in *Reviews in Computational Chemistry*, ed. Lipkowitz K. B. and Boyd D. B., Wiley-VCH, Inc., 1991, vol. 2, pp. 367–422, DOI: 10.1002/9780470125793.ch9.

22 E. Estrada, Physicochemical Interpretation of Molecular Connectivity Indices, *J. Phys. Chem. A*, 2002, **106**(39), 9085–9091, DOI: 10.1021/jp026238m.

23 H. Zhao, Q. Zhang, J. Chen, X. Xue and X. Liang, Prediction of Octanol–Air Partition Coefficients of Semivolatile Organic Compounds Based on Molecular Connectivity Index, *Chemosphere*, 2005, **59**(10), 1421–1426, DOI: 10.1016/j.chemosphere.2004.12.024.

24 J. R. Baker, J. R. Mihelcic and A. Sabljic, Reliable QSAR for Estimating Koc for Persistent Organic Pollutants: Correlation with Molecular Connectivity Indices, *Chemosphere*, 2001, **45**(2), 213–221, DOI: 10.1016/S0045-6535(00)00339-8.

25 D. L. Dowdy and T. E. McKone, Predicting Plant Uptake of Organic Chemicals from Soil or Air Using Octanol/Water and Octanol/Air Partition Ratios and a Molecular Connectivity Index, *Environ. Toxicol. Chem.*, 1997, **16**(12), 2448–2456, DOI: 10.1002/etc.5620161203.

26 A. Sabljic and M. Protić, Molecular connectivity: A novel method for prediction of bioconcentration factor of hazardous chemicals, *Chem.-Biol. Interact.*, 1982, **42**(3), 301–310, DOI: 10.1016/0009-2797(82)90074-6.

27 Q. Hu, X. Wang and M. L. Brusseau, Quantitative Structure-Activity Relationships for Evaluating the Influence of Sorbate Structure on Sorption of Organic Compounds by Soil, *Environ. Toxicol. Chem.*, 1995, **14**(7), 1133–1140, DOI: 10.1002/etc.5620140703.

28 O. G. Apul, Q. Wang, T. Shao, J. R. Rieck and T. Karanfil, Predictive Model Development for Adsorption of Aromatic Contaminants by Multi-Walled Carbon Nanotubes, *Environ. Sci. Technol.*, 2013, **47**(5), 2295–2303, DOI: 10.1021/es3001689.

29 A. Sabljić and M. Protić, Relationship between Molecular Connectivity Indices and Soil Sorption Coefficients of Polycyclic Aromatic Hydrocarbons, *Bull. Environ. Contam. Toxicol.*, 1982, **28**(2), 162–165, DOI: 10.1007/BF01608569.

30 PubChem. https://pubchem.ncbi.nlm.nih.gov/, accessed on 2025-06-18.

31 ChemDes, An integrated web-based platform for molecular descriptor and fingerprint computation, https://www.scbdd.com/chemdes, accessed on 2025-06-18.

32 S. Moavenzadeh Ghaznavi, M. Choudhary, M. Hannan, G. M. Hettiarachchi and O. G. Apul, A Critical Review of

Per- and Polyfluoroalkyl Substances Adsorption by Soil, *J. Hazard. Mater. Organics*, 2025, **1**(1), 100001, DOI: 10.1016/j.hazmo.2025.100001.

33 H. Campos-Pereira, J. Makselon, D. B. Kleja, I. Prater, I. Kögel-Knabner, L. Ahrens and J. P. Gustafsson, Binding of Per- and Polyfluoroalkyl Substances (PFASs) by Organic Soil Materials with Different Structural Composition – Charge- and Concentration-Dependent Sorption Behavior, *Chemosphere*, 2022, **297**, DOI: 10.1016/j.chemosphere.2022.134167.

34 M. Kah, D. Oliver and R. Kookana, Sequestration and Potential Release of PFAS from Spent Engineered Sorbents, *Sci. Total Environ.*, 2020, **765**, 142770, DOI: 10.1016/j.scitotenv.2020.142770.

35 F. Li, X. Fang, Z. Zhou, X. Liao, J. Zou, B. Yuan and W. Sun, Adsorption of Perfluorinated Acids onto Soils: Kinetics, Isotherms, and Influences of Soil Properties, *Sci. Total Environ.*, 2019, **649**, 504–514, DOI: 10.1016/j.scitotenv.2018.08.209.

36 H. Campos Pereira, M. Ullberg, D. B. Kleja, J. P. Gustafsson and L. Ahrens, Sorption of Perfluoroalkyl Substances (PFASs) to an Organic Soil Horizon – Effect of Cation Composition and PH, *Chemosphere*, 2018, **207**, 183–191, DOI: 10.1016/j.chemosphere.2018.05.012.

37 S. Mejia-Avendaño, Y. Zhi, B. Yan and J. Liu, Sorption of Polyfluoroalkyl Surfactants on Surface Soils: Effect of Molecular Structures, Soil Properties, and Solution Chemistry, *Environ. Sci. Technol.*, 2020, **54**(3), 1513–1521, DOI: 10.1021/acs.est.9b04989.

38 Y. Liu, F. Qi, C. Fang, R. Naidu, L. Duan, R. Dharmarajan and P. Annamalai, The Effects of Soil Properties and Co-Contaminants on Sorption of Perfluorooctane Sulfonate (PFOS) in Contrasting Soils, *Environ. Technol. Innov.*, 2020, **19**, 100965, DOI: 10.1016/j.eti.2020.100965.

39 W. Cai, D. A. Navarro, J. Du, G. Ying, B. Yang, M. J. McLaughlin and R. S. Kookana, Increasing Ionic Strength and Valency of Cations Enhance Sorption through Hydrophobic Interactions of PFAS with Soil Surfaces, *Sci. Total Environ.*, 2022, **817**, DOI: 10.1016/j.scitotenv.2022.152975.

40 K. A. Barzen-Hanson, S. E. Davis, M. Kleber and J. A. Field, Sorption of Fluorotelomer Sulfonates, Fluorotelomer Sulfonamido Betaines, and a Fluorotelomer Sulfonamido Amine in National Foam Aqueous Film-Forming Foam to Soil, *Environ. Sci. Technol.*, 2017, **51**(21), 12394–12404, DOI: 10.1021/acs.est.7b03452.

41 H. Campos-Pereira, D. B. Kleja, L. Ahrens, A. Enell, J. Kikuchi, M. Pettersson and J. P. Gustafsson, Effect of PH, Surface Charge and Soil Properties on the Solid–Solution Partitioning of Perfluoroalkyl Substances (PFASs) in a Wide Range of Temperate Soils, *Chemosphere*, 2023, **321**, DOI: 10.1016/j.chemosphere.2023.138133.

42 Y. R. Huang, S. S. Liu, J. X. Zi, S. M. Cheng, J. Li, G. G. Ying and C. E. Chen, In Situ Insight into the Availability and Desorption Kinetics of Per- and Polyfluoroalkyl Substances in Soils with Diffusive Gradients in Thin Films, *Environ. Sci. Technol.*, 2023, **57**(20), 7809–7817, DOI: 10.1021/acs.est.2c09348.

43 P. Zhou, Q. Gu, S. Zhou and X. Cui, A Novel Montmorillonite Clay-Cetylpyridinium Chloride Material for Reducing PFAS Leachability and Bioavailability from Soils, *J. Hazard. Mater.*, 2024, **465**, DOI: 10.1016/j.jhazmat.2023.133402.

44 G. Niarchos, L. Ahrens, D. B. Kleja and F. Fagerlund, Per- and Polyfluoroalkyl Substance (PFAS) Retention by Colloidal Activated Carbon (CAC) Using Dynamic Column Experiments, *Environ. Pollut.*, 2022, **308**, DOI: 10.1016/j.envpol.2022.119667.

45 D. A. Navarro, S. Kabiri, J. Ho, K. C. Bowles, G. Davis, M. J. McLaughlin and R. S. Kookana, Stabilisation of PFAS in Soils: Long-Term Effectiveness of Carbon-Based Soil Amendments, *Environ. Pollut.*, 2023, **323**, DOI: 10.1016/j.envpol.2023.121249.

46 C. E. Schaefer, D. Nguyen, E. Christie, S. Shea, C. P. Higgins and J. Field, Desorption Isotherms for Poly- and Perfluoroalkyl Substances in Soil Collected from an Aqueous Film-Forming Foam Source Area, *J. Environ. Eng.*, 2022, **148**(1), DOI: 10.1061/(asce)ee.1943-7870.0001952.

47 Q. Dong, X. Min, Y. Zhao and Y. Wang, Adsorption of Per- and Polyfluoroalkyl Substances (PFAS) by Ionic Liquid-Modified Clays: Effect of Clay Composition and PFAS Structure, *J. Colloid Interface Sci.*, 2024, **654**, 925–934, DOI: 10.1016/j.jcis.2023.10.112.

48 A. Ahmad, K. Tian, B. Tanyu and G. D. Foster, Effect of Clay Mineralogy on the Partition Coefficients of Perfluoroalkyl Substances, *ACS ES&T Water*, 2023, **3**(9), 2899–2909, DOI: 10.1021/acsestwater.3c00105.

49 J. Fabregat-Palau, A. Ershadi, M. Finkel, A. Rigol, M. Vidal and P. Grathwohl, Modeling PFAS Sorption in Soils Using Machine Learning, *Environ. Sci. Technol.*, 2025, **59**(15), 7678–7687, DOI: 10.1021/acs.est.4c13284.

50 L. Breiman, Random Forests, in *Machine Learning*, 2001, vol. 45, pp. 5–32, DOI: 10.1023/A:1010933404324.

51 R. M. O'brien, A Caution Regarding Rules of Thumb for Variance Inflation Factors, *Qual. Quantity*, 2007, **41**(5), 673–690, DOI: 10.1007/s11135-006-9018-6.

52 H. Zou and T. Hastie, Regularization and Variable Selection Via the Elastic Net, *J. Roy. Stat. Soc. B Stat. Methodol.*, 2005, **67**(2), 301–320, DOI: 10.1111/j.1467-9868.2005.00503.x.

53 K. P. Burnham, and D. R. Anderson, *Model Selection and Multimodel Inference*, ed. Burnham K. P., and Anderson D. R., Springer New York, New York, NY, 2004, DOI: 10.1007/b97636.

54 S. Lundberg, and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, 2017.

55 J.-C. Walter and G. T. Barkema, An Introduction to Monte Carlo Methods, *Phys. Stat. Mech. Appl.*, 2015, **418**, 78–87, DOI: 10.1016/j.physa.2014.06.014.

56 M. G. Arend and T. Schäfer, Statistical Power in Two-Level Models: A Tutorial Based on Monte Carlo Simulation, *Psychol. Methods*, 2019, **24**(1), 1–19, DOI: 10.1037/met0000195.

57 G. Derringer and R. Suich, Simultaneous Optimization of Several Response Variables, *J. Qual. Technol.*, 1980, **12**(4), 214–219, DOI: 10.1080/00224065.1980.11980968.

58 J. Xie, S. Liu, L. Su, X. Zhao, Y. Wang and F. Tan, Elucidating Per- and Polyfluoroalkyl Substances (PFASs) Soil-Water Partitioning Behavior through Explainable Machine Learning Models, *Sci. Total Environ.*, 2024, **954**, DOI: 10.1016/j.scitotenv.2024.176575.

59 A. C. Umeh, M. Hassan, M. Egbuatu, Z. Zeng, M. Al Amin, C. Samarasinghe and R. Naidu, Multicomponent PFAS Sorption and Desorption in Common Commercial Adsorbents: Kinetics, Isotherm, Adsorbent Dose, PH, and Index Ion and Ionic Strength Effects, *Sci. Total Environ.*, 2023, **904**, 166568, DOI: 10.1016/J.SCITOTENV.2023.166568.

60 O. G. Apul, Y. Zhou and T. Karanfil, Mechanisms and Modeling of Halogenated Aliphatic Contaminant Adsorption by Carbon Nanotubes, *J. Hazard. Mater.*, 2015, **295**, 138–144, DOI: 10.1016/j.jhazmat.2015.04.030.

61 X. Fu, J. Sun, K. Tian, Y. Liu and H. Zhang, Predicting the Sorption Capacity of Perfluoroalkyl and Polyfluoroalkyl Substances in Soils: Meta-Analysis and Machine Learning Modeling, *Environ. Sci. Technol.*, 2025, **59**(33), 17699–17710, DOI: 10.1021/acs.est.4c11313.

62 J. Xie, S. Liu, L. Su, X. Zhao, Y. Wang and F. Tan, Elucidating Per- and Polyfluoroalkyl Substances (PFASs) Soil-Water Partitioning Behavior through Explainable Machine Learning Models, *Sci. Total Environ.*, 2024, **954**, 176575, DOI: 10.1016/j.scitotenv.2024.176575.

63 C. P. Higgins and R. G. Luthy, Sorption of Perfluorinated Surfactants on Sediments, *Environ. Sci. Technol.*, 2006, **40**(23), 7251–7256, DOI: 10.1021/es061000n.

64 L. Ahrens, L. W. Y. Yeung, S. Taniyasu, P. K. S. Lam and N. Yamashita, Partitioning of Perfluorooctanoate (PFOA), Perfluorooctane Sulfonate (PFOS) and Perfluorooctane Sulfonamide (PFOSA) between Water and Sediment, *Chemosphere*, 2011, **85**(5), 731–737, DOI: 10.1016/j.chemosphere.2011.06.046.

65 H. Chen, M. Reinhard, V. T. Nguyen and K. Y.-H. Gin, Reversible and Irreversible Sorption of Perfluorinated Compounds (PFCs) by Sediments of an Urban Reservoir, *Chemosphere*, 2016, **144**, 1747–1753, DOI: 10.1016/j.chemosphere.2015.10.055.

66 C. Yin, C.-G. Pan, S.-K. Xiao, Q. Wu, H.-M. Tan and K. Yu, Insights into the Effects of Salinity on the Sorption and Desorption of Legacy and Emerging Per-and Polyfluoroalkyl Substances (PFASs) on Marine Sediments, *Environ. Pollut.*, 2022, **300**, 118957, DOI: 10.1016/j.envpol.2022.118957.

67 N. H. Batjes, World Soil Property Estimates for Broad-Scale Modelling (WISE30sec), https://www.isric.org/documents/document-type/isric-report-201501-world-soil-property-estimates-broad-scale-modelling, 2015, Wageningen, accessed on 2025-06-18.

68 C. Abou-Khalil, D. Sarkar, P. Braykaa and M. C. Boufadel, Mobilization of Per- and Polyfluoroalkyl Substances (PFAS) in Soils: A Review, *Curr. Pollut. Rep.*, 2022, **8**(4), 422–444, DOI: 10.1007/s40726-022-00241-8.

