Machine Learning Frameworks for Predicting Contaminant Leaching and Sorption Dynamics in Environmental Matrices

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Amirhossein Ershadi
aus Rasht/Iran

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 10.10.2025

Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Peter Grathwohl
2. Berichterstatter:	Prof. Dr.-Ing. Wolfgang Nowak
3. Berichterstatter:	Prof. Dr. James Craig

ABSTRACT

Machine learning (ML) has the potential to fundamentally transform environmental science by providing a rapid, scalable alternative to slow, resource-intensive experiments and complex simulations. Traditional experimental methods, such as column leaching tests, require weeks of continuous laboratory measurements. Similarly, numerical contaminant transport models demand substantial computational resources to simulate complex subsurface processes accurately. These time and resource requirements limit the practicality of such approaches for large-scale environmental assessments. This thesis introduces ML as a transformative tool to rapidly and accurately predict contaminant behavior.

ML-driven models significantly reduce the duration of column leaching tests from 7–14 days to a single day by leveraging early-stage experimental data. Trained on construction and demolition waste materials, these models not only accurately predict the long-term leaching behavior of key contaminants—including sulfate, vanadium, chromium, copper, and organic pollutants (15 priority PAHs identified by the US-EPA)—but also enable fast, cost-effective risk assessments and guide sustainable waste management decisions. Once the ML models are trained, sensitivity analysis is performed to understand the underlying leaching dynamics, revealing pH and electrical conductivity as the most influential factors.

Beyond experimental acceleration, this work enhances the efficiency of numerical simulations used to model contaminant transport in porous media. Traditional numerical models rely on solving coupled equations for advection, dispersion, and inter-phase mass transfer, requiring extensive computational resources. To address this, a surrogate modeling framework is developed using a random forest stacking model, reducing computational costs by over 1,000 times while maintaining predictive accuracy. Optimized through adaptive-recursive sampling, the model efficiently selects training data points, balancing exploration and exploitation. Additionally, Neural Posterior Estimation calibrates model parameters probabilistically using copper leaching data from two distinct soils, enabling robust uncertainty quantification.

Furthermore, the need for accurate predictive tools is particularly urgent for emerging contaminants such as per- and polyfluoroalkyl substances (PFAS), known for their persistence, and complex sorption behavior. This thesis develops the PFAS Sorption Stacking Model (PSSM)—an ML-driven framework integrating compound-specific properties with soil characteristics to predict solid-liquid distribution coefficients (K_d). A key innovation is the incorporation of charge density as a sorption descriptor, particularly for PFAS compounds with *pKa* values near typical soil pH, which provides deeper insights into electrostatic interactions. A comprehensive dataset consolidating sorption isotherm data for 51 PFAS compounds across 455 soil types enhances model robustness. Missing values are systematically imputed using a k-nearest neighbors algorithm, preserving data integrity and improving predictive stability. The model demonstrates accurate and stable predictive performance, achieving a normalized root mean square error of 0.07, making it one of the most precise PFAS sorption models to date. As a practical application, PSSM is integrated into an online platform for real-time PFAS sorption predictions and the generation of spatial K_d maps. This tool supports targeted contamination assessments, facilitating the identification of PFAS hotspots and improving environmental risk management.

ZUSAMMENFASSUNG

Maschinelles Lernen (ML) hat das potential, Umweltwissenschften grundlegend zu verändern, indem es eine schnelle, skalierbare Alternative zu langwierigen, ressourcenintensiven Experimenten und komplexen Simulationen bietet. Traditionelle experimentelle Methoden wie Säulenelutionsversuche erfordern mehrere Wochen kontinuierlicher Laboruntersuchungen. Vergleichbar benötigen numerische Modelle zum Schadstofftransport erhebliche Rechenressourcen, um komplexe Prozesse im in Böden und Grundwasserleitern präzise zu simulieren. Dieser Zeit- und Ressourcenaufwand konventioneller Ansätze schränkt deren Anwendbarkeit für großräumige Umweltuntersuchungen ein. In dieser Arbeit wird ML als innovativ Werkzeug vorgestellt, das Vorhersagen zum Schadstoffverhalten mit hoher Geschwindigkeit und Genauigkeit beschleunigt.

ML-basierte Modelle verkürzen die Dauer von Säulenelutionsversuchen von 7 bis 14 Tagen auf nur einen Tag, indem sie Daten aus frühen Phasen eines Experimentes nutzen. Auf Bau- und Abbruchabfällen trainiert, sagen diese Modelle das langfristige Auslaugungsverhalten relevanter Schadstoffe—inklusive Sulfat, Schwermetalle (Vanadium, Chrom, Kupfer) und organische Schadstoffe (15 prioritäre PAK gemäß US-EPA)— mit hoher Genauigkeit voraus. Darüber hinaus ermöglichen der Ansatz auch schnellere, kostengünstigere Risikobewertungen und unterstützen nachhaltige Entscheidungen im Abfallmanagement. Nach dem Training der ML-Modelle wird eine Sensitivitätsanalyse durchgeführt, um die zugrunde liegenden Elutionsdynamiken besser zu verstehen. Dabei erweisen sich der pH-Wert und die elektrische Leitfähigkeit als die einflussreichsten Parameter.

Über die Verkürzung von Experimenten hinaus steigert diese Arbeit auch die Effizienz numerischer Simulationen, die zum Modellieren des Schadstofftransports in porösen Medien eingesetzt werden. Traditionelle numerische Modelle beruhen auf der Lösung gekoppelter Gleichungen für Advektion, Dispersion und Stoffübertragung durch intrapartikuläre Porendiffusion, die entsprechend hohe Rechenleistung erfordert. Um dem zu begegnen, wurde ein Ersatzmodellierungs-Framework auf Basis eines Random-Forest-Stacking-Modells entwickelt. Somit konnte die Rechenkosten um mehr als das Tausendfache reduziert werden, ohne an Vorhersagegenauigkeit einzubüßen. Optimiert durch adaptiv-rekursives Sampling wählt das Modell effizient Trainingsdatenpunkte aus und findet ein Gleichgewicht zwischen Exploration und Ausnutzung. Darüber hinaus kalibriert die Neural Posterior Estimation die Modellparameter probabilistisch mithilfe von Kupferelutionsdaten aus zwei unterschiedlichen Böden, was eine robuste Quantifizierung von Unsicherheiten ermöglicht.

Der Bedarf an präzisen Vorhersagetools ist besonders dringlich bei neuartigen Schadstoffen wie per- und polyfluorierten Alkylsubstanzen (PFAS), die für ihre Persistenz, und komplexen Sorptionsmechanismen bekannt sind. In dieser Arbeit wird das PFAS Sorption Stacking Model (PSSM) vorgestellt—ein ML-basiertes Framework, das substanzspezifische Eigenschaften und Bodeneigenschaften einbezieht, um Fest-Flüssig-Verteilungskoeffizienten (K_d) vorherzusagen. Eine wesentliche Neuerung ist die Einbeziehung der Ladungsdichte als Sorptionsdeskriptor, insbesondere bei PFAS-Verbindungen mit Säurekonstanten (pKa-Werten) nahe dem typischen Boden-pH. Dies ermöglicht tiefere Einblicke in elektrostatische Wechselwirkungen. Ein umfassender Daten-

satz vereint Sorptionsisothermen für 51 PFAS-Verbindungen in 455 Bodentypen und erhöht so die Robustheit des Modells. Fehlende Werte werden systematisch über einen Nächste-Nachbarn-Klassifikation imputiert, wodurch die Datenintegrität erhalten bleibt und die Vorhersagestabilität steigt. Das Modell liefert präzise und robuste Vorhersagen mit einem normalisierten Quadratwurzel des mittleren quadratischen Fehlers von 0,07 und zählt damit zu den genauesten bisher verfügbaren PFAS-Sorptionsmodellen. Um diese Erkenntnisse in die Praxis zu überführen, wird PSSM in eine Online-Plattform integriert, die in Echtzeit PFAS-Sorptionsprognosen sowie räumliche K_d-Karten erstellt. Dieses Werkzeug ermöglicht zielgerichtete Kontaminations-analysen, erleichtert die Identifizierung von PFAS-Hotspots und verbessert das Umwelt-Risikomanagement.

The only way we can change is to be real with ourselves.
If you want to be great, you have to go through real suffering.
You can't achieve anything worthwhile if you're afraid of the challenge.
Embrace the pain, and you'll find the strength within.

— David Goggins

ACKNOWLEDGEMENTS

The journey to completing this dissertation has been filled with challenges, learning experiences, and moments of growth. Along the way, I have been fortunate to have the support of incredible people whose guidance, encouragement, and kindness made all the difference. Their contributions, whether big or small, shaped both this work and my personal development. I am deeply grateful to each and every one of them.

My Supervisors

Prof. Dr. Peter Grathwohl and Dr. Michael Finkel – Your guidance, support, and encouragement have been invaluable throughout this journey. Your expertise shaped the direction of my research, and your mentorship made the challenges easier to navigate. I couldn't have asked for better supervisors. Thank you for everything!

Advisors

Prof. Dr. Olaf Cirpka – Your insightful advice and constructive feedback greatly enhanced the quality of my work. I truly appreciate the time and effort you dedicated to helping me refine my ideas.

Dr. Bernd Susset – Thank you for generously providing the column leaching test data, which was a crucial foundation for this research.

The Amazing GUZ Administration & IT Teams

Dr. Wolfgang Bott, Dr. Peter Merkel, Artur, Iris, Eva, Patricia, Marion, Mirjam, and Monika – You kept everything running smoothly, and I can't thank you enough for your patience and efficiency.

Uli, Marko, and Philipp – thank you for your technical support. A special thanks to Willi Kappler for always having my back with server issues. I truly appreciate your help and support.

Colleagues & Friends Who Made This Journey Special

Joel, Dominik, Hermann, Anton, Touraj, Shuya, Baharat, and Alex – Thank you for the countless discussions, ideas, and support along the way. Working alongside you made this journey more insightful and enjoyable.

Ran and Binlong – My awesome officemates! Your camaraderie and encouragement made long days in the office much more bearable. I'm grateful for the learning, the laughs, and the shared experiences.

Kleio, Philipp, Marie Madeleine, Akash, James, Caroline, Al, Vjerjan, Guy, Sara, and Daniel – thank you for your support, encouragement, and shared moments. I truly appreciate each one of you.

Family

Finally, to my brother, Reza (*Dadash*[1]) – Your constant support, belief in me, and words of encouragement have been my source of strength and inspiration. I dedicate this dissertation to you.

To my family – Your unconditional love and encouragement have been my foundation, every step of the way. None of this would have been possible without you.

1 In Farsi, Dadash means older brother

CONTENTS

1 INTRODUCTION 1
1.1 Solid Waste Reuse and Environmental Challenges . . . 1
1.2 Leaching Tests: Evaluating Contaminant Release and Mobility . . . 2
1.3 Mechanisms Governing Contaminant Transport in Column . . . 4
1.4 Machine Learning: A New Paradigm for Contaminant Mobility Assessment 6
1.4.1 Data Preprocessing . . . 6
1.4.2 Learning Algorithms . . . 7
1.4.3 Model Optimization and Validation . . . 9
1.4.4 Surrogate Models . . . 10
1.5 Parameter Estimation . . . 11
1.5.1 Bayesian Inference . . . 11
1.5.2 Simulation-Based Inference . . . 12
1.6 Sensitivity Analysis . . . 12
1.7 Aim and Objectives . . . 13
1.8 Thesis Structure . . . 13
2 APPLICABILITY OF MACHINE LEARNING MODELS FOR THE ASSESSMENT OF LONG-TERM POLLUTANT LEACHING FROM SOLID WASTE MATERIALS 15
2.1 Introduction . . . 16
2.2 Materials and Methods . . . 17
2.2.1 Leaching Test . . . 17
2.2.2 Compilation of Available Data . . . 18
2.2.3 Data Preparation . . . 19
2.2.4 Machine Learning Models . . . 20
2.2.5 Regression Analysis . . . 21
2.2.6 Hyperparameters Optimization: Cross-Validation . . . 22
2.2.7 Feature Importance . . . 23
2.2.8 Model Performance . . . 24
2.3 Results and Discussion . . . 24
2.3.1 Overall Performance of the Models at LS=2 and 4 . . . 24
2.3.2 Overall Performance of the Models at LS=10 . . . 26
2.3.3 Influence of the Size of the Training data set on Model Performance 28
2.3.4 Key Model Features . . . 28
2.4 Summary and Conclusion . . . 30
3 ENSEMBLE SURROGATE MODELING OF ADVECTIVE-DISPERSIVE TRANSPORT WITH INTRAPARTICLE DIFFUSION MODEL FOR COLUMN LEACHING TEST 33
3.1 Introduction . . . 34
3.2 Theory and Background . . . 37
3.2.1 Process-Based Column Leaching Model . . . 37
3.2.2 Surrogate Models . . . 39
3.2.3 Ensemble Model . . . 46
3.2.4 Model Calibration . . . 47
3.3 Results and Discussion . . . 48

3.3.1 Verification of the Surrogate Model . . . 48
3.3.2 Analysis of Posterior Distributions . . . 51
3.4 Conclusions . . . 53
4 MODELING PFAS SORPTION IN SOILS USING MACHINE LEARNING 55
4.1 Introduction . . . 56
4.2 Materials and Methods . . . 57
4.2.1 PFAS considered in this work . . . 57
4.2.2 Criteria for data compilation . . . 59
4.2.3 Model Development . . . 60
4.3 Results and Discussion . . . 64
4.3.1 Effect of chain length in PFAS sorption . . . 64
4.3.2 Effect of functional group in PFAS sorption . . . 64
4.3.3 Identification of outliers and anomalous trends . . . 65
4.3.4 Model Performance and Sensitivity Analysis . . . 66
4.4 Environmental Implications . . . 69
5 CONCLUSION&OUTLOOK 71
5.1 Synthesis of Major Findings . . . 71
5.1.1 How does machine learning transform leaching test assessments? . 71
5.1.2 How are computational bottlenecks in leaching simulations overcome? . . . 71
5.1.3 What new insights are provided into PFAS sorption behavior? . . . 72
5.2 Future Prospects . . . 73
5.2.1 Expanding Datasets for Enhanced Predictive Performance . . . 73
5.2.2 Broadening the Scope to Emerging Contaminants . . . 73
5.2.3 Applying Sorption Models on a Global Scale . . . 73
5.2.4 Application of Sorption Coefficients in Contaminant Transport Models . . . 74
5.2.5 Extending Surrogate Models for Broader Environmental Applications 74
5.2.6 Towards Integrated Environmental Management Platforms . . . 74
A APPENDIX FOR CHAPTER 1 75
A.1 Overview of 15 Priority PAHs by US-EPA . . . 75
A.2 Column leaching test standards . . . 76
A.3 Missing data imputation . . . 76
A.3.1 Basic imputation techniques . . . 76
A.3.2 *k*-nearest neighbors (KNN) . . . 77
A.4 Data scaling . . . 77
A.5 Unsupervised and Reinforcement Learning . . . 78
A.6 Linear Regression . . . 78
A.7 Ensemble algorithms . . . 79
A.7.1 Random Forest . . . 79
A.7.2 Extremely Randomized Trees (ExtraTrees) . . . 79
A.7.3 Gradient Boosting . . . 79
A.7.4 Extreme Gradient Boosting (XGBoost) . . . 80
A.8 Training process of multi-layer perceptrons (MLP) . . . 80
A.8.1 Core training steps . . . 81
A.8.2 Regularization techniques . . . 81
A.9 Model performance evaluation metrics . . . 82

B APPENDIX FOR CHAPTER 2 85
B.1 Relevant compounds and the threshold concentration in aqueous leachate according to German recycling decree . . . 85
B.2 Selected range and tuned hyperparameters of algorithms . . . 85
B.3 Overall performance of sequence-timepoint, hybrid and early-stage input models to predict concentration at LS=10 . . . 86
B.4 Distribution of the pH, Electrical Conductivity, and DOC in the data set . . 86
B.5 Schematic of random forest algorithm . . . 87
B.6 Model validation: R^2 distribution . . . 87
B.7 Cumulative concentrations predicted by ML models . . . 88
B.8 Programming environment and computer system used to develop and analyse the models . . . 88
B.9 Bootstrapping . . . 88
B.10 Random-Search method . . . 89
C APPENDIX FOR CHAPTER 3 91
C.1 Empirical models for K_d values of heavy metals in soil . . . 91
C.2 Distribution coefficients K_d of organic compounds . . . 91
C.3 Pairwise projections of the virtual reality datasets (VRD) . . . 93
C.4 Correlation coefficient analysis for posterior parameter distribution . . . 94
C.5 Range of applicability of developed ensemble surrogate model for selected heavy metals . . . 95
C.6 Range of applicability of developed ensemble surrogate model for organic compounds . . . 96
C.7 Constant parameters for ADE-IPD model (DIN 19528) . . . 96
C.8 Impact of surrogate model combination order on stacking model performance 97
D APPENDIX FOR CHAPTER 4 99
D.1 : pH-speciation diagram for some representative PFAS species . . . 99
D.2 : Additional information on the derivation of K_d data from literature studies 115
D.3 : Number of soils and entries used to derive log K_{OC} values . . . 117
D.4 : Construction of a ML soil property imputer model based on KNN . . . 118
D.5 : Statistical tests used to assess log K_{OC} distributions across different PFAS subfamilies . . . 121
D.6 : Comparison of model performance with other available tools . . . 128
D.7 : Additional geospatial K_d (PFAS) maps . . . 131

BIBLIOGRAPHY 137

LIST OF FIGURES

Figure 1.1 Batch test setup . . . 2
Figure 1.2 Column leaching test setup . . . 3
Figure 1.3 Mass transfer limited by film diffusion and intraparticle diffusion 5
Figure 1.4 Decision tree framework . . . 7
Figure 1.5 Stacking ensemble framework . . . 8
Figure 1.6 Schematic of a Multi-Layer Perceptron . . . 9
Figure 1.7 Simulation-Based Inference framework . . . 12
Figure 2.1 Data Distributions of Leachate Concentrations (Half-Violin Plots) 19
Figure 2.2 Flowchart of Compound-Specific Multi-Output ML Model Development and Assessment . . . 21
Figure 2.3 Repeated Five-Fold Cross-Validation chart . . . 23
Figure 2.4 Measured vs. Predicted Concentrations at LS=2 and LS=4 for Test Set Samples . . . 25
Figure 2.5 Long-Term Leaching Behavior: Measured vs. ML Predictions . . . 27
Figure 2.6 Impact of Training Data Size on ML Model Performance . . . 28
Figure 2.7 SHAP Beeswarm Plot of Key Feature Importance . . . 29
Figure 3.1 Spatial Distribution of Virtual Reality Datasets and Training Sets 36
Figure 3.2 Column Leaching Test Flowchart (German Standard DIN 19528) . 37
Figure 3.3 Numerical Model Setup for 1-D Transport with Intraparticle Diffusion . . . 39
Figure 3.4 Flowchart of Surrogate Modeling for ADE-IPD Transport . . . 40
Figure 3.5 Optimization of Surrogate Models via Adaptive Sampling . . . 49
Figure 3.6 Surrogate Model Performance for Cumulative Concentrations (RF Stacking & IDW) . . . 51
Figure 3.7 Model Calibration Comparison for Copper in Soils A and B using NPE . . . 52
Figure 4.1 Relative distribution of PFAS subfamilies and compounds in the K_d (PFAS) dataset . . . 59
Figure 4.2 Trend between log K_{OC} and log K_{OW} for PFAS species . . . 66
Figure 4.3 Comparison of predicted and measured log K_d values . . . 68
Figure 4.4 Sensitivity analysis: SHAP and Partial Dependence for PFAS K_d model . . . 69
Figure 4.5 Predicted log K_d values for PFOSB in Europe . . . 70
Figure B.1 Violin Plots of pH, EC, and DOC Distributions Across LS Ranges . 86
Figure B.2 Schematic of Random Forest Averaging Method . . . 87
Figure B.3 R^2 Scores from Repeated K-Fold CV (LS=2 & 4) . . . 87
Figure B.4 Cumulative Concentrations at LS=2: Measured vs. Predicted . . . 88
Figure C.1 Parameter Space Coverage: VRDs and DoE Sampling . . . 93
Figure C.2 Correlation Coefficients Between Posterior Parameters for Soils A and B . . . 94
Figure C.3 Heatmaps of Ensemble Surrogate Model Applicability (pH, %OC, and K_d for Metals) . . . 95

Figure C.4 Heatmap of K_{OC} vs. Organic Carbon Content and K_d 96
Figure D.1 Spatial distribution of the soil samples used to develop the KNN model 118
Figure D.2 Soil and PFAS properties ranges 119
Figure D.3 Entry distribution over organic content ranges for some selected PFAS species 119
Figure D.4 Visualization of soil pH, C_{org} and texture information for training and validation sets 120
Figure D.5 Statistical test flowchart 121
Figure D.6 Distribution of K_{OC} data for different PFCA 122
Figure D.7 Distribution of K_{OC} data for different PFSA 123
Figure D.8 Distribution of K_{OC} data for different FOSA 123
Figure D.9 Distribution of K_{OC} data for different FTOH 124
Figure D.10 Distribution of K_{OC} data for different FTS 124
Figure D.11 Distribution of K_{OC} data for different PFPA 125
Figure D.12 Distribution of K_{OC} data for different Betaine 125
Figure D.13 Distribution of K_{OC} data for different PFPiA 126
Figure D.14 Distribution of K_{OC} data for different PFAS species with four fluorinated carbons 126
Figure D.15 Distribution of K_{OC} data for different PFAS species with six fluorinated carbons 127
Figure D.16 Distribution of K_{OC} data for different PFAS species with eight fluorinated carbons 127
Figure D.17 Distribution of K_{OC} data for different PFAS species with ten fluorinated carbons 128
Figure D.18 Visualization of the comparative prediction accuracy of available K_d prediction tools 130
Figure D.19 Spatial distribution of selected soil properties within the LUCAS soil repository 131
Figure D.20 Geospatial K_d (TFA) information 132
Figure D.21 Geospatial K_d (PFOA) information 133
Figure D.22 Geospatial K_d (PFOS) information 134
Figure D.23 Geospatial K_d (PFOSB) information 135

LIST OF TABLES

Table 2.1 Overall performance of each model to predict actual concentration at LS=2 and LS=4 using the best-tuned algorithms. 25
Table 3.1 Prior Parameter Distribution of ADE-IPD Model (DIN 19528) . . . 42
Table A.1 The 15 (US-EPA) PAHs along with their molecular formulas and structures 75
Table A.2 Overview of column leaching test standards 76
Table B.1 The "threshold" concentrations of the examined organic and inorganic substances for various classes of CDW (Bundesgesetzblatt, 2021). 85
Table B.2 Tuned Hyperparameters for LS=2 and 4 predictive ML models. . . 85
Table B.3 Summary of key metrics for each model. 86
Table C.1 Constant Parameters for ADE-IPD Model (DIN 19528) 96
Table C.2 Surrogate Models Required for Stacking (RRMSE $\leq$ 10%) 97
Table D.1 Summary of PFAS Entries and log K_{OC} 117
Table D.2 Reported metrics evaluating the performance for different K_d (PFAS) prediction tools 129

STATEMENT OF CONTRIBUTIONS

This thesis is based on three co-authored manuscripts, two published and one submitted to the Environmental Science & Technology. As the first author of each paper, I was primarily responsible for the design of the models and the analysis, contributing significantly to the core findings and outcomes.

ORCiD: 0000-0001-6691-0056
Google Scholar: Amirhossein Ershadi

First Author

1. **Published in WasteManagement**, September 2023
   Ershadi, A., Finkel, M., Susset, B., Grathwohl, P.: Applicability of machine learning models for the assessment of long-term pollutant leaching from solid waste materials, *WasteManagement*.
   https://doi.org/10.1016/j.wasman.2023.09.001, 2023.

2. **Published in Journal of Contaminant Hydrology**, November 2024,
   Ershadi, A., Finkel, M., Liu, B., Cirpka, O., Grathwohl, P.: Ensemble surrogate modeling of advective-dispersive transport with intraparticle diffusion model for column-leaching test, *Journal of Contaminant Hydrology*.
   https://doi.org/10.1016/j.jconhyd.2024.104423, 2024.

3. **Published in Environmental Science & Technology**, April 2025,
   Ershadi, A., Fabregat-Palau, J., Finkel, M., Rigol, A., Vidal, M., Grathwohl, P.: Modeling PFAS Sorption in Soils Using Machine Learning, *Environmental Science & Technology*.
   https://pubs.acs.org/doi/10.1021/acs.est.4c13284, 2025.

1 INTRODUCTION

1.1 SOLID WASTE REUSE AND ENVIRONMENTAL CHALLENGES

Globally, billions of tons of solid waste—such as demolition waste, industrial byproducts, and agricultural residues—are generated each year, with Germany alone accounting for approximately 340 million tonnes (Shah et al., 2021; BMUV, 2023). Reusing these materials has become a cornerstone of sustainable resource management by reducing reliance on virgin resources, diverting waste from landfills, and advancing circular economy objectives (European Commission, 2018). However, the environmental challenges associated with material reuse are complex, as embedded contaminants can leach into soil and water systems, posing significant risks to ecosystems and human health (Butera et al., 2015).

Construction and demolition waste (CDW) exemplifies both the potential and challenges of material reuse. As a key contributor to recycling initiatives, CDW includes materials such as concrete, bricks, and wood, which are widely repurposed in infrastructure projects, thereby reducing the need for virgin resource extraction (Dahlbo et al., 2015; BMUV, 2023). However, the presence of contaminants in CDW, including sulfates, heavy metals, and polycyclic aromatic hydrocarbons (PAHs)—specifically the 15 priority PAHs identified by the US-EPA (listed in Table A.1)—necessitates an evaluation of their environmental impact and long-term sustainability implications (Susset & Grathwohl, 2011). Sulfates, leaching from gypsum-based materials, can elevate salinity in aquatic systems, disrupting water quality and ecosystems (Jang & Townsend, 2001). Heavy metals such as vanadium, chromium, and copper—originating from concrete additives, wiring, paints, and treated wood—pose diverse environmental and health risks. Vanadium bioaccumulates, disrupting metabolic processes and causing oxidative stress and kidney damage (Barceloux & Barceloux, 1999; Rehder, 2015). Chromium, especially in its hexavalent form (Cr(VI)), is highly soluble under alkaline conditions, posing carcinogenic risks and reducing soil microbial diversity (Prado et al., 2016; Shahid et al., 2017; Sharma et al., 2022). Copper, although essential in trace amounts, disrupts soil enzymatic activity and contributes to soil degradation at elevated levels (Bernard et al., 2009; Alengebawy et al., 2021). Furthermore, PAHs originating from asphalt and tar-based materials are persistent and hydrophobic, posing long-term exposure risks, with compounds like benzo[a]pyrene classified as carcinogenic (Johnsen et al., 2005; Butera et al., 2015; Zhang et al., 2023).

Building on these challenges, emerging contaminants like per- and polyfluoroalkyl substances (PFAS)[1] present additional threats to environmental sustainability due to their persistence, ubiquity, and associated risks. Often referred to as *forever chemicals*, PFAS resist chemical, biological, and thermal degradation, due to the strength of the carbon-fluorine bond (Prevedouros et al., 2006; Buck et al., 2011). Since their introduction in the 1950s, PFAS have been widely used in industrial and consumer products—including non-stick cookware, stain-resistant textiles, firefighting foams, and grease-resistant packag-

1 Per- and polyfluoroalkyl substances: "Per" refers to compounds where all hydrogens in the carbon chain are replaced by fluorine atoms, while "poly" refers to compounds with multiple but not all hydrogens replaced by fluorine.

ing—offering functional benefits while raising complex environmental and human health concerns (Glüge et al., 2020). The widespread use of PFAS has led to their presence in soils, sediments, the atmosphere, and water systems worldwide (Cousins et al., 2020; Brusseau et al., 2020). Long-chain PFAS, such as perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS), bioaccumulate in ecosystems, binding strongly to organic matter and persisting in the environment for decades. In contrast, short-chain PFAS, like trifluoroacetic acid (TFA), exhibit high mobility in aqueous systems due to their greater solubility and limited sorption to soil particles, thereby increasing the risk of groundwater and surface water contamination (Nguyen et al., 2020).

1.2 LEACHING TESTS: EVALUATING CONTAMINANT RELEASE AND MOBILITY

Understanding how contaminants are released from solid materials into the environment is essential for effectively managing their impact and mitigating associated ecological concerns. Leaching tests are essential tools in environmental science, simulating water-solid interactions to quantify how contaminants dissolve or desorb into seepage water, groundwater, or surface waters (Grathwohl & Sloot, 2007; Imoto, 2024). These tests serve as valuable tools for evaluating the mobility and persistence of both organic and inorganic pollutants under controlled conditions (Kalbe et al., 2007). Various leaching test methodologies—batch tests, column (or percolation) tests, and lysimeter tests—are employed to address the complexities of contaminant behavior. Each test offers unique advantages, complementing one another to provide a comprehensive assessment of contaminant release under varying environmental conditions.

Batch tests, illustrated in Figure 1.1, are recognized for their simplicity and efficiency in assessing contaminant release. They were initially developed to evaluate the leaching potential of contaminants from sewage sludge (DIN 38414, 1984). These tests involve mixing a predefined mass of material with an extraction fluid at a controlled liquid-to-solid (LS) ratio [$L^3\,M^{-1}$], typically ranging from 2:1 to 10:1 by weight, under standardized laboratory conditions. The mixture is mechanically agitated, typically for 24 to 72 hours, to ensure an equilibrium state between the solid and liquid phases. Following agitation, the liquid phase is separated for analysis, commonly using filtration or centrifugation techniques.

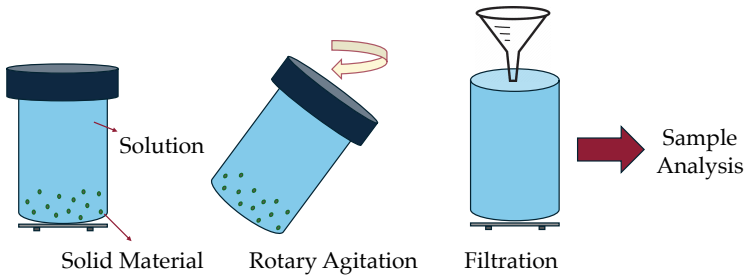


Figure 1.1: Schematic diagram of batch test setup.

While batch tests offer rapid assessments, their primary limitation lies in providing a single result at a fixed LS ratio, which fails to capture the dynamic nature of real-world conditions (Naka et al., 2016). The LS ratios commonly applied in batch tests are often much larger than those observed in the field (e.g., 0.25 L kg^{-1}), resulting in significantly lower aqueous concentrations compared to field conditions due to the dilution effect, particularly for compounds with low solid-liquid distribution coefficients (K_d [$L^3\,M^{-1}$]):

$$C_w = \frac{C_{s,\text{ini}}}{K_d + LS} \tag{1.1}$$

where C_w is the aqueous concentration of the contaminant in the batch test [$\mathrm{M\,L^{-3}}$], and $C_{s,\text{ini}}$ is the initial contaminant concentration in the solid phase [$\mathrm{M\,M^{-1}}$]. Additionally, experimental artifacts, such as the mobilization of fine particles, the formation of emulsions, and potential cross-contamination during filtration or centrifugation, can further compromise the accuracy and reliability of results (Grathwohl & Susset, 2009).

To address these limitations, column leaching tests (depicted in Figure 1.2) simulate the flow of water through solid material closer to natural conditions, providing a dynamic and controlled approach for studying contaminant mobility and release (Grathwohl & Susset, 2009). These tests involve packing the material into a cylindrical column and percolating water through it, typically from the bottom to ensure even flow, avoid entrapped air, and minimize preferential pathways. Effluent samples are collected systematically over time and analyzed to generate leaching curves that describe contaminant release as a function of LS ratio or time. The protocols for column tests are governed by internationally established standards, such as those outlined in Table A.2, which ensure consistency in parameters like column dimensions, flow rates, and particle size distributions. While these tests provide valuable insights into the leaching of potential pollutants, they require specialized equipment, precise material preparation, and extended durations (e.g., 14 days) to capture time-dependent behavior, making them resource-intensive in terms of time and labor.

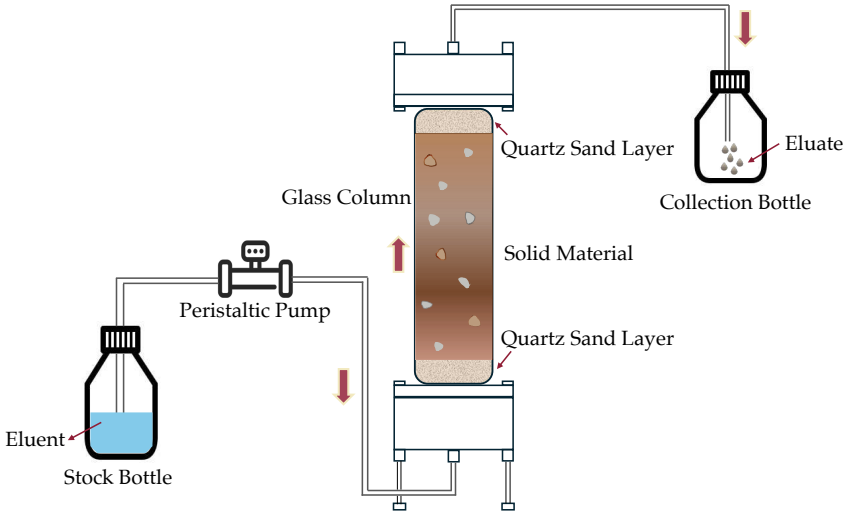


Figure 1.2: Schematic diagram of column leaching test setup.

Lysimeter studies extend laboratory-scale tests by simulating natural conditions, such as precipitation-driven infiltration, over prolonged periods. These studies consider dynamic factors like microbial activity, seasonal variations, and soil heterogeneity, providing valuable insights into the long-term behavior of contaminants. By bridging laboratory experiments with field-scale observations, lysimeter tests support the calibration and validation of transport models while enhancing the understanding of real-world contaminant mobility. However, their implementation demands substantial resources, time, and maintenance (Corwin, 2000; Dąbrowska et al., 2019; Asadollahi et al., 2020).

1.3 MECHANISMS GOVERNING CONTAMINANT TRANSPORT IN COLUMN

The movement of contaminants through porous media in column leaching tests is often described by the Advection-Dispersion Equation (ADE), which captures two key processes: *advection*, representing the bulk flow of solutes with water, and *dispersion*, accounting for spreading caused by variations in flow velocities and pathways (Grathwohl & Susset, 2009; Liu et al., 2021). The general form of the one-dimensional (1-D) governing equation ADE is given as

$$\frac{\partial}{\partial t}\left(nC_w + \rho_b C_s\right) + \frac{\partial}{\partial x}\left(nvC_w - nD_L\frac{\partial C_w}{\partial x}\right) = 0, \tag{1.2}$$

where C_w is the solute concentration in water [$\mathrm{M\,L^{-3}}$], C_s is the concentration in the solid phase [$\mathrm{M\,M^{-1}}$], ρ_b is the bulk density [$\mathrm{M\,L^{-3}}$], t is time [T], x is the length of the column [L], n is the intergranular porosity [−], v is the seepage velocity of the water [$\mathrm{L\,T^{-1}}$], and D_L is the longitudinal dispersion coefficient [$\mathrm{L^2\,T^{-1}}$].

For local equilibrium conditions, the concentration in the solid phase (C_s) is in equilibrium with the solute concentration in water (C_w) through the solid-liquid distribution coefficients K_d [$\mathrm{L^3\,M^{-1}}$] ($= \frac{C_s}{C_w}$). Substituting this relationship into the ADE, the equation 1.2 simplifies under equilibrium assumptions to:

$$\frac{\partial C_w}{\partial t} = \frac{D_L}{R}\frac{\partial^2 C_w}{\partial x^2} - \frac{v}{R}\frac{\partial C_w}{\partial x}, \tag{1.3}$$

where R represents the retardation factor [−] and is defined as:

$$R = 1 + \frac{\rho_b K_d}{n}. \tag{1.4}$$

Despite this simplification, real-world systems often deviate from equilibrium due to mass transfer limitations, which arise when the rate of solute exchange between the solid and liquid phases is slower than the timescale of transport (Grathwohl, 2014). These limitations can significantly impact contaminant migration, particularly in porous media with heterogeneities in grain size, porosity, or surface properties (Liu et al., 2021). Incorporating mass transfer effects into transport models provides a more realistic representation of solute behavior, especially in dynamic systems such as column leaching tests, and is particularly important for compounds with slow desorption rates, where long-term leaching is primarily governed by desorption kinetics.

Two key mechanisms contribute to mass transfer limitations, as illustrated in Figure 1.3: *film diffusion* and *intraparticle pore diffusion* (Grathwohl, 2014; Finkel & Grathwohl, 2017; Liu et al., 2021, 2022a). Mass transfer of contaminants between particles and the bulk solution may be limited by diffusion from the particle surface through a stagnant aqueous boundary layer, a process known as external film diffusion. This process is commonly described by the linear driving force model, which expresses the mass transfer rate as proportional to the concentration gradient across the boundary layer:

$$\frac{\partial C_w}{\partial t} = kA_o(C_w' - C_w), \tag{1.5}$$

where C_w' is the solute concentration at the solid-water interface [$\mathrm{M\,L^{-3}}$], k is the mass transfer coefficient [$\mathrm{L\,T^{-1}}$], $A_o = \frac{6m_d}{V_w\rho_s d}$ is the specific surface area of particles (i.e., spherical)

per unit volume of water in the column [L^{-1}], m_d is the dry mass of the solids [M], d is the particle diameter [L], V_w is the water volume [L^3], and ρ_s is the particle density [$M\,L^{-3}$]. The mass transfer coefficient k can be related to the film thickness δ_f [L], which represents the stagnant boundary layer of water through which solutes must diffuse. The film thickness can be estimated using the dimensionless Sherwood number (Sh):

$$\delta_f = \frac{d}{Sh}. \tag{1.6}$$

In contrast, intraparticle pore diffusion involves the transport of solutes within the porous structure of solid particles, which is governed by Fick's second law (e.g., for spherical particles):

$$\frac{\partial C_{w,\text{intra}}}{\partial t} = D_a \left(\frac{\partial^2 C_{w,\text{intra}}}{\partial r^2} + \frac{2}{r} \frac{\partial C_{w,\text{intra}}}{\partial r} \right), \tag{1.7}$$

where $C_{w,\text{intra}}$ is the solute concentration in intraparticle pore water [$M\,L^{-3}$], r is the radial distance [L], and D_a is the apparent diffusion coefficient [$L^2\,T^{-1}$]. The apparent diffusion coefficient, D_a, is expressed as:

$$D_a = \frac{D_e}{\epsilon + K_d \rho_p}, \tag{1.8}$$

where ϵ is the intraparticle porosity [−], D_e is the effective diffusion coefficient [$L^2\,T^{-1}$], and ρ_p is the bulk density of the porous particles [$M\,L^{-3}$]. Solute release by pore diffusion leads to prolonged tailing effects due to increasing diffusion distances over time.

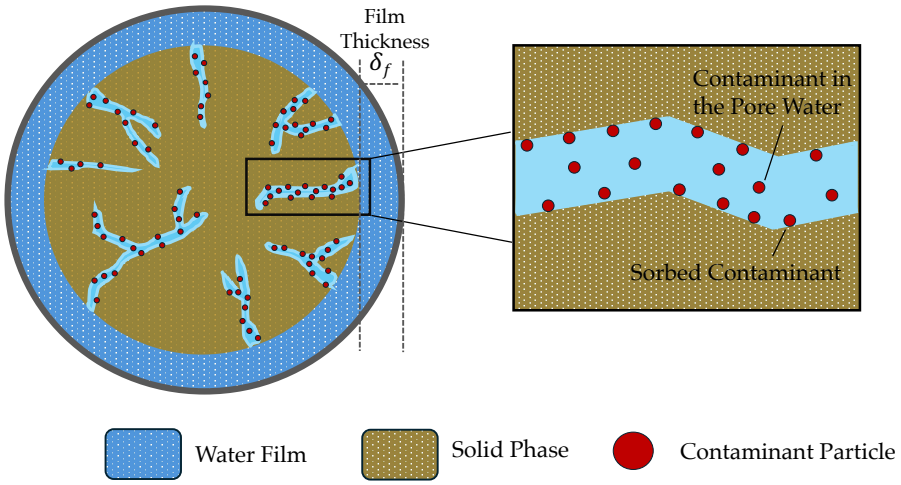


Figure 1.3: Scheme of mass transfer limited by film diffusion (left) and intraparticle diffusion (right).

Both mechanisms—film diffusion and intraparticle diffusion—are essential for accurately modeling contaminant migration in porous media. Film diffusion primarily dominates for compounds with large K_d, especially in the early stages of transport when solutes rapidly interact with the solid phase. In contrast, intraparticle diffusion becomes significant for compounds with smaller K_d values and limited intraparticle porosity, as these conditions facilitate a shift in mass transfer resistance from the external film to the intraparticle pore space, resulting in slower, diffusion-controlled transport within the porous matrix (Liu et al., 2021).

1.4 MACHINE LEARNING: A NEW PARADIGM FOR CONTAMINANT MOBILITY ASSESSMENT

Classification, while fundamental to many ML applications, is not explored in this dissertation as it falls outside the scope of the presented work.

The increasing complexity of contaminant mobility in the environmental systems necessitates efficient and scalable approaches for predicting and evaluating their transport dynamics, fate, and broader environmental impact. Traditional experimental and numerical methods, while foundational, are often time-intensive and computationally demanding, limiting their applicability for large-scale or dynamic scenarios. Recent advancements in data-driven techniques, particularly machine learning (ML), provide transformative tools that complement and enhance these conventional methods. By leveraging the growing availability of experimental datasets and computational power, ML facilitates the extraction of meaningful patterns and relationships, enabling accurate and rapid predictions for complex systems.

Machine learning fundamentally seeks to approximate an unknown function f that maps input features X to a target output Y, represented as:

$$Y = f(X) + \varepsilon, \tag{1.9}$$

where ε accounts for noise or uncertainty in the system (Russell & Norvig, 2022). Models learn underlying patterns in the data and generalize them to unseen scenarios by optimizing a loss function tailored to the specific objectives of regression or classification.

1.4.1 *Data Preprocessing*

Data preprocessing is essential for robust ML workflows, addressing missing data, feature scaling, and data splitting. Imputation methods handle missing values, with basic approaches like mean or median substitution offering simplicity but potentially skewing data distributions. Advanced techniques, such as *K-Nearest Neighbors (KNN)* and *Random Forest Imputation*, leverage data patterns to predict missing entries more accurately, effectively capturing non-linear relationships and dependencies (Troyanskaya et al., 2001; Tang & Ishwaran, 2017). For additional details on data imputation, refer to Appendix A.3.

Standardization and scaling adjust features to ensure consistent contributions, particularly in models sensitive to magnitude differences. *Standardization* (Z-score scaling) centers data to a mean of zero and a variance of one, while *Min-Max scaling* normalizes values to a fixed range (e.g., [0, 1]), enhancing model performance and stability (see Appendix A.4). Equally important is proper data splitting, which ensures reliable model evaluation by partitioning datasets into training and test (i.e., unseen data) sets, often in proportions such as 80%-20%. The *holdout method* is a widely used approach for this purpose, dividing the dataset into subsets to facilitate unbiased model assessment. For datasets without inherent structure, *random splitting* is commonly employed, where the data is randomly divided into subsets, ensuring that each instance has an equal chance of being included in the training or test set. For imbalanced datasets, *stratified splitting* is employed to preserve class distributions across subsets, maintaining representativeness. In time-dependent datasets, *chronological splitting* ensures that the temporal order is respected by using earlier data for training and later data for testing, preventing data leakage[2].

2 Data leakage occurs when information from the test set or future data is inadvertently used during training, leading to overly optimistic model performance.

1.4.2 *Learning Algorithms*

This dissertation primarily focuses on supervised learning, while unsupervised and reinforcement learning are briefly touched upon in Appendix A.5

At the core of machine learning, supervised learning provides the foundation for modeling relationships between input features and target variables using labeled datasets. It enables models to generalize from known examples to unseen data, making it an essential paradigm for predictive and decision-making tasks.

Linear regression, a foundational technique, models the linear relationship between input features and the target output. While straightforward and interpretable, it is prone to overfitting, especially in high-dimensional settings where the number of features is large relative to the number of observations. Overfitting in linear regression typically occurs when the model captures noise or irrelevant patterns in the data, especially if polynomial or interaction terms are introduced to model nonlinear relationships. To reduce overfitting and enhance generalization, regularization techniques such as *Ridge regression* and *Lasso regression* extend the linear regression framework by introducing penalty terms into the loss function. Ridge regression adds an L_2-norm penalty to the loss function, which minimizes the sum of squared coefficients, effectively shrinking coefficients of uninformative predictors toward zero. In contrast, Lasso regression incorporates an L_1-norm penalty, which minimizes the sum of the absolute values of the coefficients. This property of the L_1-norm not only reduces coefficient magnitudes but also forces some coefficients to become exactly zero, leading to sparsity in the model (Tibshirani, 1996; Hastie, 2020). The detailed mathematical formulations of these methods are presented in Appendix A.6.

For non-linear systems, tree-based methods such as *Decision Trees* offer high interpretability, making them valuable tools for regression tasks (see Figure 1.4). These models determine optimal splits by minimizing a loss function, evaluating all possible divisions, and selecting the one that minimizes the total error across child nodes (see Eq. 1.10). However, despite their clarity and simplicity, decision trees are highly prone to overfitting, particularly when grown to excessive depths (Schaffer, 1993).

$$\mathcal{L}_{\text{split}} = \sum_{i \in L} \mathcal{L}(Y_i - \bar{Y}_L)^2 + \sum_{j \in R} \mathcal{L}(Y_j - \bar{Y}_R)^2, \tag{1.10}$$

where $\mathcal{L}$ is the loss function, L and R represent the left and right child nodes, Y_i and Y_j are the target values in those nodes, and $\bar{Y}_L$ and $\bar{Y}_R$ are the mean target values in the respective subsets. The process continues recursively until a stopping criterion, such as a maximum tree depth or minimum samples per leaf, is met (Bertsimas & Dunn, 2017).

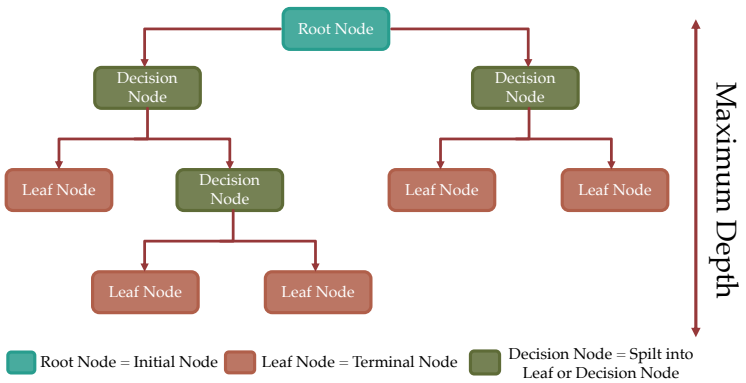


Figure 1.4: Schematic of a decision tree, illustrating the partitioning process with decision rules at internal nodes and predictions at leaf nodes.

To address the limitations of single decision trees, *ensemble techniques* combine multiple models to enhance robustness and predictive accuracy. Among these, *Random Forests* construct a collection of diverse decision trees by utilizing bootstrapped[3] datasets and random feature selection. The aggregated predictions of these trees reduce variance and improve stability, making Random Forests a powerful alternative to standalone decision trees (Breiman, 2001). Building on the idea of randomness, *Extremely Randomized Trees (Extra Trees)* introduce additional randomness by selecting feature splits at random thresholds. This approach not only improves computational efficiency but also enhances resilience to noise, making Extra Trees well-suited for handling noisy datasets (Geurts et al., 2006). Another widely used ensemble method, *Gradient Boosting*, takes a different approach by building trees sequentially. Each subsequent tree aims to correct the residual errors from the previous iteration, gradually improving overall model performance (Friedman, 2001). To further optimize this process, *Extreme Gradient Boosting (XGBoost)* incorporates regularization, parallelization, and sparsity-aware algorithms, offering greater scalability and precision for large datasets (Chen & Guestrin, 2016). Detailed explanations and mathematical formulations of each method are provided in Appendix A.7.

In addition to these ensemble strategies, *stacking models*, also known as stacked generalization, offers a powerful approach to enhance predictive performance (Wolpert, 1992). Unlike methods such as boosting, which iteratively correct errors by focusing on difficult samples, stacking employs a second layer of learning (i.e., meta-model) to learn the optimal combination of predictions from multiple base-models (Shams et al., 2021; Ershadi et al., 2024). This strategy capitalizes on the strengths of diverse algorithms while mitigating their individual weaknesses, resulting in a robust predictive framework (see Figure 1.5).

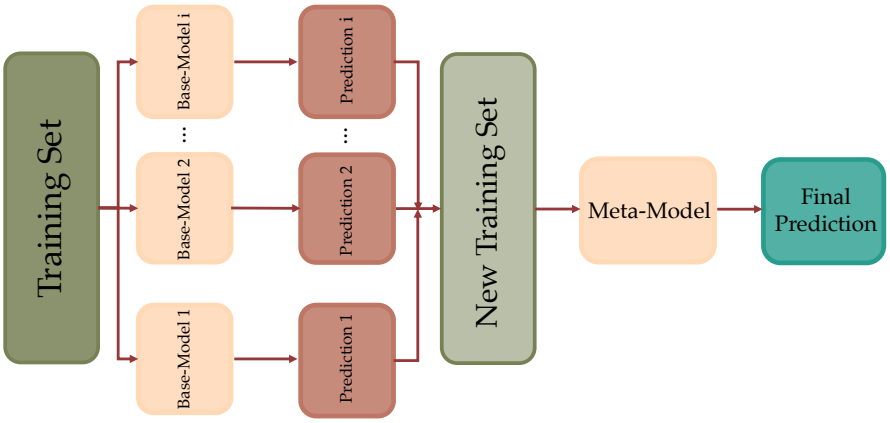


Figure 1.5: Visualization of a stacking ensemble framework, where base models generate predictions that are combined by a meta-model to produce the final output.

3 Bootstrapping involves randomly sampling data with replacement to create subsets for training each tree.

For capturing more complex, non-linear patterns, *Multi-Layer Perceptrons (MLPs)* are a widely used neural network architecture. An MLP comprises interconnected layers of neurons: an input layer, one or more hidden layers, and an output layer. Each layer performs linear transformations followed by non-linear activation functions, enabling the network to model intricate relationships within the data (see Figure 1.6). The output of an MLP with a single hidden layer can be represented as:

$$\hat{Y} = \sigma\left(W_2 \cdot \sigma(W_1 X + b_1) + b_2\right), \tag{1.11}$$

where W_1 and W_2 are weight matrices, b_1 and b_2 are biases, and σ is a non-linear activation function such as *Rectified Linear Unit (ReLU)*, *Hyperbolic Tangent (Tanh)*, or *sigmoid*. The non-linearity introduced by σ allows MLPs to capture intricate patterns that linear models cannot (Goodfellow et al., 2016). Training MLPs involves optimizing hyperparameters like the number of layers, neurons per layer, activation function, learning rate, and regularization techniques (e.g., dropout, weight decay) to balance model complexity and generalization (Dubey et al., 2022). A detailed explanation of the training process for MLP is provided in Appendix A.8.

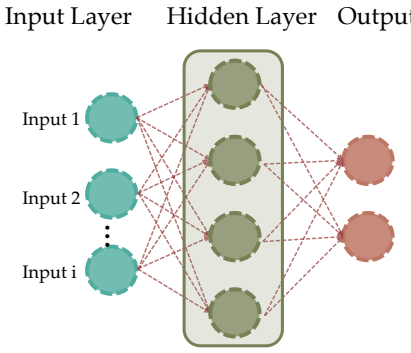


Figure 1.6: Schematic of a Multi-Layer Perceptron (MLP) architecture, showing input, hidden, and output layers with non-linear transformations.

1.4.3 *Model Optimization and Validation*

Optimization and validation are pivotal in constructing reliable ML models, ensuring both peak performance and robust generalization to unseen data. Central to this process is the distinction between model *parameters* and *hyperparameters*, each playing a unique role in shaping the model's behavior and performance. Model parameters, such as weights in neural networks or coefficients in regression models, are learned directly from the training data during the learning process by iteratively minimizing a loss function. However, the efficiency and success of this internal learning process are heavily influenced by hyperparameters—user-defined settings that govern how the model learns.

Hyperparameters define the structure and dynamics of the learning process, ranging from learning rates and batch sizes to the complexity of the model architecture, such as the number of layers in a neural network or the depth of decision trees. These settings are not derived directly from the data but instead must be tuned to create the optimal conditions for learning. Poorly chosen hyperparameters can lead to underfitting, where the model fails to capture essential data patterns, or overfitting, where it learns noise rather than meaningful trends.

Identifying the best hyperparameter configuration is critical and requires systematic optimization techniques. While methods like *grid search* explore a comprehensive range of

predefined hyperparameter combinations, they can be computationally intensive (Ogunsanya et al., 2023). More efficient approaches, such as *random search*, balance thoroughness with practicality by sampling configurations randomly across the hyperparameter space (Bergstra & Bengio, 2012). Advanced techniques like *Bayesian Optimization* further enhance efficiency by leveraging probabilistic models to focus on the most promising regions of the hyperparameter space, reducing computational cost while improving the likelihood of finding optimal configurations (Feurer & Hutter, 2019).

Validation assesses a model's ability to generalize to unseen data, ensuring robust and reliable predictions. *Cross-validation (CV)* is a widely employed method for model evaluation, offering a comprehensive approach to assess performance. The most common variant, *k-fold cross-validation*, partitions the dataset into k subsets or folds. The model is iteratively trained on $k-1$ folds and tested on the remaining fold, cycling through all folds to compute average performance metrics. This reduces the variability inherent in a single train-test split and provides a reliable measure of generalization. *Repeated k-fold cross-validation* further enhances this process by performing the k-fold procedure multiple times with different random splits, improving robustness. For small datasets, *Leave-One-Out Cross-Validation (LOOCV)* provides high reliability by using each data point as a test set in turn, ensuring minimal bias. However, its computational demands increase significantly with larger datasets (Wong, 2015). For imbalanced datasets, *stratified cross-validation* preserves class proportions across folds, ensuring fair assessments and reducing bias (Allgaier & Pryss, 2024). Quantitative evaluation of model performance often involves metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Relative Root Mean Squared Error (RRMSE), Ratio of Performance to Deviation (RPD), and the Determination Coefficient (R^2). A detailed description of each metric, including its mathematical formulation and interpretation, is provided in Appendix A.9.

1.4.4 *Surrogate Models*

Surrogate models, also known as proxy models, are computationally efficient approximations designed to replicate the behavior of complex and resource-intensive systems (Razavi et al., 2012). By capturing the input-output relationships of a target function $f(X)$, these models significantly reduce the need for high-fidelity simulations, preserving accuracy while lowering computational costs (Allgeier & Cirpka, 2023).

A critical step in constructing surrogate models is determining how training samples are distributed across the input space $\mathcal{X}$. This process, known as the Design of Experiments (DoE), directly influences the surrogate's ability to approximate the true function $f(X)$. Structured sampling methods, such as Latin Hypercube Sampling and quasi-random sequences (e.g., Sobol and Halton methods), ensure uniform coverage of the design space, optimizing resource use and enabling incremental refinement in adaptive workflows (Halton, 1960; Sobol', 1967; Sheikholeslami & Razavi, 2017; Garud et al., 2017). Once training data is established, function approximation methods construct the surrogate $\hat{f}(X)$. Techniques range from simple polynomial regression for low-dimensional problems to Gaussian Process Regression (GPR) for non-linear systems and Radial Basis Functions (RBF) for localized variations. Advanced approaches, such as neural networks and ensemble methods, expand the capabilities of surrogate models to handle highly complex or high-dimensional problems (Shams et al., 2021; Ershadi et al., 2024).

Surrogate models can be implemented in offline or adaptive-recursive frameworks, each suited to specific scenarios. Offline frameworks rely on static surrogates constructed from a fixed dataset generated through DoE (Razavi et al., 2012). These models are computationally efficient when the initial dataset sufficiently represents the design space. However, their static nature limits adaptability in capturing localized complexities or addressing under-sampled regions, particularly in high-dimensional or multi-modal spaces.

In contrast, adaptive-recursive frameworks iteratively refine the surrogate model during the optimization process by incorporating new data points. This dynamic approach balances exploration—reducing uncertainty in under-sampled regions—and exploitation—focusing on regions predicted to yield the most promising outcomes (Jones et al., 1998; Zou et al., 2007). By continuously updating the surrogate model, these frameworks enhance predictive reliability, making them particularly well-suited for complex tasks such as sensitivity analysis and parameter estimation.

1.5 PARAMETER ESTIMATION

Parameter estimation, also referred to as model calibration, is a process focused on optimizing model parameters to achieve the best agreement between simulated and observed data. The calibration process aims to identify the optimal parameter set Θ^* within the parameter space Θ, minimizing the discrepancy between model outputs $M(\Theta)$ and observed data Y_{obs}, often expressed as:

$$\Theta^* = \arg\min_{\Theta} L(Y_{\text{sim}}, Y_{\text{obs}}). \tag{1.12}$$

Here, $Y_{\text{sim}} = M(\Theta)$ represents simulated outputs, and L is a loss function. Calibration methods range from deterministic optimization (e.g., gradient descent (Prasad et al., 2020), Levenberg-Marquardt (Moré, 1978; Kleefeld & Reißel, 2011) to probabilistic approaches (e.g., Bayesian inference (Von Toussaint, 2011; Van De Schoot et al., 2021), Markov Chain Monte Carlo (Vrugt, 2016)), depending on system complexity and analysis goals.

1.5.1 *Bayesian Inference*

Bayesian inference provides a probabilistic framework for model calibration, combining prior knowledge of the parameters with observational data to estimate their posterior distributions (Oladyshkin & Nowak, 2019). This approach is governed by Bayes' theorem:

$$p(\Theta \mid Y_{\text{obs}}) = \frac{p(Y_{\text{obs}} \mid \Theta)p(\Theta)}{p(Y_{\text{obs}})}, \tag{1.13}$$

where $p(\Theta \mid Y_{\text{obs}})$ is the posterior distribution, $p(Y_{\text{obs}} \mid \Theta)$ is the likelihood, and $p(\Theta)$ is the prior distribution. The posterior $p(\Theta \mid Y_{\text{obs}})$ encapsulates both the updated beliefs about the parameters after observing the data and the uncertainty associated with these estimates.

The likelihood function $p(Y_{\text{obs}} \mid \Theta)$ measures how well the model with parameters Θ explains the observed data. When combined with the prior $p(\Theta)$, which reflects existing knowledge or assumptions about the parameters, the posterior provides a comprehensive characterization of the parameter space.

1.5.2 *Simulation-Based Inference*

For scenarios where the likelihood function $p(Y_{\text{obs}} \mid \Theta)$ is intractable, Simulation-Based Inference (SBI) offers a robust likelihood-free approach by using forward simulations to approximate posterior distributions (Cranmer et al., 2020). Among SBI methods, Neural Posterior Estimation (NPE) has emerged as a highly efficient and scalable approach, particularly in high-dimensional and computationally expensive settings (Papamakarios & Murray, 2016; Lueckmann et al., 2017). NPE directly learns the posterior distribution $p(\Theta \mid Y_{\text{obs}})$ from simulated data using neural density estimators, such as normalizing flows (Papamakarios et al., 2019). Normalizing flows transform a simple base distribution (e.g., a Gaussian) into a complex posterior through a series of invertible, parameterized transformations, enabling the modeling of intricate posterior structures (Kobyzev et al., 2021).

By leveraging amortized inference, which concentrates computational effort during training, NPE enables rapid and efficient posterior evaluations for subsequent observations, making it particularly suitable for iterative workflows and real-time applications (Greenberg et al., 2019). This efficiency allows NPE to capture complex parameter-data relationships while significantly reducing computational demands compared to traditional inference methods. Figure 1.7 illustrates the framework of SBI.

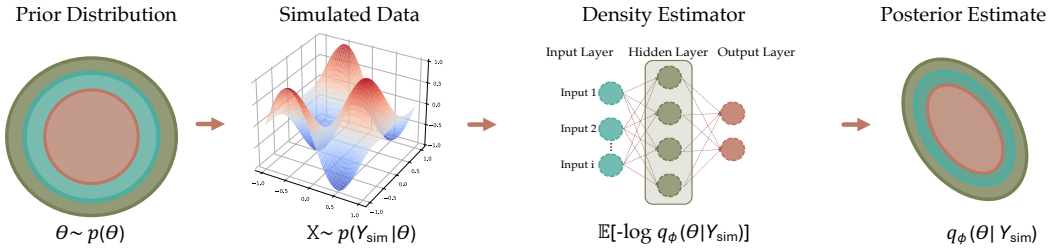


Figure 1.7: Simulation-Based Inference (SBI) framework for Bayesian parameter estimation.

1.6 SENSITIVITY ANALYSIS

Understanding a model's sensitivity to input features is essential for interpreting predictions and assessing reliability. Two primary methods for sensitivity analysis are SHapley Additive exPlanations (SHAP) and Partial Dependence (PD). SHAP, grounded in cooperative game theory, provides model-agnostic and locally accurate explanations by quantifying each feature's contribution to a prediction (Shapley, 1953). It achieves this by evaluating the average marginal impact of a feature across all possible combinations of feature subsets (Lundberg & Lee, 2017). This ensures that each feature's importance is fairly and consistently attributed, regardless of the model's complexity or structure. Partial Dependence, in contrast, provides a global view by analyzing the average marginal effect of a specific feature on the predicted outcomes across the dataset (Friedman, 2001; Greenwell et al., 2018). By averaging out the influence of other features, PD highlights general trends and the direction of feature impacts, offering an intuitive understanding of the relationship between input features and predictions.

1.7 AIM AND OBJECTIVES

Understanding and predicting the behavior of contaminants in environmental systems remain a critical scientific and practical challenge. Contaminants interact with soils and reused materials in complex ways, governed by a multitude of factors including chemical properties, soil composition, and environmental conditions. Traditional approaches, such as column leaching tests and numerical models, have laid the groundwork for our understanding but are often constrained by their labor-intensive nature and significant computational demands. These limitations restrict their applicability to large-scale or dynamic scenarios, highlighting the need for innovative and scalable methodologies.

Emerging computational techniques, combined with increasingly comprehensive experimental datasets, offer a unique opportunity to transform the evaluation of contaminant leaching and sorption processes. This thesis aims to bridge the gap between traditional methods and data-driven approaches by integrating machine learning, surrogate modeling, parameter estimation, and sensitivity analysis.

1.8 THESIS STRUCTURE

CHAPTER 2 This chapter introduces machine learning models developed to predict the leaching behavior of contaminants from construction and demolition waste materials, addressing the limitations of time-intensive column leaching tests. By leveraging measurements at lower liquid-to-solid (LS) ratios, these models estimate leachate concentrations at higher LS values, offering a faster and more efficient alternative to traditional methods. The chapter outlines the development and validation of these models, highlighting their application for predicting long-term leaching behavior. Sensitivity analysis identifies pH, electrical conductivity, and compound-specific concentrations as critical factors influencing predictions. The results demonstrate how machine learning can enhance leaching assessments, supporting sustainable material reuse and environmental protection. Supporting information for this chapter is provided in Appendix B.

CHAPTER 3 This chapter develops ensemble surrogate models to address the computational challenges of simulating column leaching tests, which evaluate the leaching behavior and environmental risks of contaminated soils and waste materials. These tests are governed by complex solute transport and kinetic inter-phase mass transfer dynamics, traditionally modeled using the advective-dispersive framework. However, when sorption kinetics involve intraparticle diffusion, the modeling process becomes significantly more complex, requiring fine discretization along both the column flow axis and within grain interiors. This additional computational burden makes inverse modeling and sensitivity analysis particularly demanding. To overcome these limitations, two surrogate modeling approaches were developed: random forest stacking and inverse-distance weighted interpolation. The models were optimized using adaptive sampling techniques that balance exploration and exploitation of the parameter space, guided by infill criteria such as expected improvement and Mahalanobis distance. Integrated with Neural Posterior Estimation, these models enabled robust inference of parameter distributions and provided accurate predictions of copper leaching behavior from two distinct soils, significantly reducing computational demands. Supporting information for this chapter is detailed in Appendix C.

CHAPTER 4 This chapter addresses the challenges of predicting the sorption behavior of per- and polyfluoroalkyl substances (PFAS) in soils, a key factor in assessing their environmental mobility and risks. The diverse properties of PFAS, combined with their complex interactions with soil-specific parameters, make accurate predictions difficult. To overcome these challenges, PFAS Sorption Stacking model (PSSM), a novel machine learning model, was developed to predict the solid-liquid distribution coefficients (K_d) of a wide range of PFAS. Built on a comprehensive dataset of 51 PFAS compounds across 455 soil and sediment samples, the model integrates molecular weight, hydrophobicity, and speciation (pK_a) with soil properties such as pH, organic carbon content, and texture. The results highlight the dominant role of hydrophobic interactions, exemplified by octanol-water partition coefficient (K_{ow}), over electrostatic factors like pK_a and cation exchange capacity in influencing sorption processes. By combining these predictions with soil repository data, PSSM generates K_d maps, providing a scalable and practical framework for environmental risk assessment. Supporting methodologies and additional analyses are detailed in Appendix D.

CHAPTER 5 This chapter summarizes the key findings, emphasizing the development of efficient machine learning tools for modeling contaminant leaching and sorption. The advancements demonstrated in this work highlight the potential for scalable environmental risk assessments. The chapter concludes with an outlook on future research directions, including extending datasets for column leaching tests and PFAS, modeling the sorption behavior of emerging contaminants like antibiotics, and integrating advanced modeling frameworks to deepen the understanding of contaminant transport processes.

2

APPLICABILITY OF MACHINE LEARNING MODELS FOR THE ASSESSMENT OF LONG-TERM POLLUTANT LEACHING FROM SOLID WASTE MATERIALS

Amirhossein Ershadi[1], **Michael Finkel**[1], **Bernd Susset**[1], **Peter Grathwohl**[1]

[1]*Department of Geosciences, University of Tübingen, Schnarrenbergstraße 94-96, 72076 Tübingen, Germany*

Published in the journal of **WasteManagement**
https://doi.org/10.1016/j.wasman.2023.09.001

Author Contributions

First author:
Scientific ideas: 70%, Data processing: 90%,
Analysis & interpretations: 70%, Paper writing: 80%

2.1 INTRODUCTION

The increase in production and agglomeration of solid waste and its management have caused environmental problems all over the world. Most of the solid waste comes from human activities such as demolition and renovation of buildings, steel production, waste, and coal incineration (European Commission, 2014).

Construction and Demolition Waste (CDW), consisting of concrete, natural aggregates, bricks, plaster, and occasionally wood, glass, and plastics, constitutes the largest fraction (approximately 30%) of the solid waste produced in the European Union (Rodrigues et al., 2013; European Commission, 2014). While being considered for a long time as unusable waste and being disposed in landfills, in the meantime, recycled CDW is known to be technically feasible for various construction applications (Del Rey et al., 2015). At present, over 90% of CDW is reused in technical structures such as road constructions, embankments, concrete, dams, etc. (Molenaar & van Niekerk, 2002; Rahman et al., 2014; Lupsea et al., 2014; Rahman et al., 2015; Kreislaufwirtschaft Bau, 2018). This high level of reuse remarkably reduces the need for quarrying virgin materials and lowers environmental risks from waste disposal. The material itself, but also the presence of adhesives, paints, sealants, PCBs, asbestos, and lead-based paint, may contain both organic and inorganic contaminants (Del Rey et al., 2015), and pose a potential risk of subsoil and groundwater contamination, owing to the release of pollutants if water percolates through these materials (Susset & Grathwohl, 2011). Thus, appropriate measures and testing methods are required to ensure the safe reuse of CDW in the environment.

Leaching tests provide a means to determine the aqueous concentrations at which contaminants are released from various sources, including soils (Kalbe et al., 2014; Naka et al., 2016; Löv et al., 2019; Röhler et al., 2021), recycled construction materials (Butera et al., 2015; Del Rey et al., 2015; Bandow et al., 2019; López-Uceda et al., 2019; Diotti et al., 2020; Prieto-Espinoza et al., 2022), radioactive or other waste materials (Hyks et al., 2009; Chen et al., 2021). These tests also allow to evaluate the cumulative mass release based on the liquid-to-solid ratio (LS), where the LS represents the ratio between the amount of liquid, i.e., water (L), and the amount of solid (S) in the test expressed in L kg^{-1} dry matter.

Several methods of leaching tests with different initial conditions have been developed to evaluate the release of pollutants from porous materials in the laboratory: column tests (DIN 19528, 2009; US EPA Method 1314, 2013; DIN CEN/TS 16637-2, 2014), batch tests (DIN 38414-6, 1986; DIN EN 12457-1, 2003; DIN 19529, 2015), and sequential leaching tests (NEN 7349, 1995; Comans & Roskam, 2002). Table A.2 presents an overview of some column leaching standards.

Column leaching tests are preferred over batch tests since they can simulate the water flow through porous materials closer to natural conditions, have better reproducibility than batch and sequential leaching tests, and can assess the leaching behavior of the pollutants over extended periods (Roussat et al., 2008; Grathwohl & Susset, 2009).

In preparation for the German decree for the use of recycled mineral materials (Bundesgesetzblatt, 2021), the leaching potential of CDWs (amongst other materials) was evaluated based on a series of column leaching tests with material samples from diverse origins. Among a larger number of identified constituents, pollutants of concern were determined by a statistical evaluation of cumulative concentrations at LS=2 and predictions of pollutant release and leachate concentrations in real-world reuse applications in comparison to existing quality standards. According to Susset, Grathwohl, 2011, five compounds can

be found in the leachate at levels of environmental concern: Sulfate, 15 (US-EPA) PAHs, Chromium, Copper, and Vanadium. The eluate concentration of these compounds may exceed the quality standard for certain reuse applications. For the practical implementation of the decree, and to regulate the reusability, CDWs are divided into three different categories (CDW-1, CDW-2, CDW-3) of material quality in terms of contaminant levels in the leachate from low to high (Table B.1).

Artificial Intelligence (AI) and Machine Learning (ML) algorithms have increasingly been applied in the field of chemistry and geochemistry due to their ability to find patterns, model dynamic systems, and make predictions about complex processes. Substantial research papers have relied on ML models using Artificial Neural Network (ANN), and tree-based algorithms to predict the leaching behavior of chemical compounds from soils and solid wastes (Bayar et al., 2009; Bazoobandi et al., 2019; Flores et al., 2020; Lillington et al., 2020; Spijker et al., 2021; Zhang et al., 2022). ML methods have also been employed to investigate the bioavailability of pollutants in water and their removal processes from solids (Wu et al., 2013; Caglar Gencosman & Eker Sanli, 2021).

In this study, we implement parametric and non-parametric ML algorithms to quantify the leaching of chemical compounds from CDW materials, recognizing the complexity of the governing processes and material properties. We investigate whether the use of ML models that are tailored to effectively predict long-term leaching behavior using limited short-term measurement data predictions can replace long-term measurements. This would shorten the duration of extended column leaching tests, which typically require a significant duration, depending on the specific test setup in the lab. Such tests typically take 3 to 4 days (> 7 days) to reach LS=4 (LS=10) (Grathwohl & Susset, 2009).

To this end, we (i) build an exhaustive data set of the leaching behavior of chemical compounds by collecting the results of experimental studies from the last decade, (ii) develop several ML models to predict the long-term leaching behavior of CDWs at LS=2, 4, 10, and in addition (iii) analyze the impact of individual chemical features and the size of the data set on the performance of the proposed ML models.

2.2 MATERIALS AND METHODS

2.2.1 *Leaching Test*

Leaching tests are conducted to evaluate the potential risk of harmful pollutants present in recycled waste materials being released into seepage water and ultimately contaminating groundwater (Grathwohl & Sloot, 2007; Kalbe et al., 2014). These tests evaluate the leachability of the contaminants rather than quantifying their total content in the solid phase.

The leaching tests take into consideration several crucial factors to determine the leachability of contaminants. These factors, including chemical compound properties such as solubility and distribution coefficients, as well as material characteristics like porosity, particle size, permeability, composition, and organic carbon content, along with physical parameters such as flow rate, temperature, contact time, pH, redox conditions and LS ratio, are considered in leaching tests (van der Sloot, 2004; Kalbe et al., 2008; Liu et al., 2021).

Generally, leaching tests are assessed in terms of the cumulative release of mass for specific values of the **L**-to-**S** ratio. L/S ratio or LS represents the time after a specific volume of liquid (i.e., water) percolated through the column filled with a specific mass of solids. The

eluate, that is, the solution obtained from the leaching experiment, is typically collected according to the predefined separate fractions of LS (Kalbe et al., 2007). Results obtained in the laboratory can be mapped to the field, i.e., real-world leaching scenarios via LS, provided that local equilibrium between solid and water exists. The time scale in the field that is associated with a certain LS value can be estimated by the following equation:

$$T_{\text{field}} = 2.65 \cdot (1 - n) \cdot d/q \cdot LS, \tag{2.1}$$

where n, d, and q denote the porosity [−], the thickness of the release zone [L], and percolation rate [L T^{-1}], respectively (Finkel & Grathwohl, 2017). Note that, for typical bulk densities and seepage water rates, T_{field} ranges for LS = 2 between 10 and 20 years.

An "extended column leaching test" is used to characterize the long-term leaching behavior of pollutants at the different LS ratios (e.g., 0.3, 1, 2, 4, and - optionally - 10). These long-term measurements can provide researchers with valuable information about the release processes and effectiveness of soil or waste management systems over time, enabling them to ensure long-term sustainability and reduce the risk of environmental contamination.

For the practical implementation of risk assessment by means of column leaching tests, typically the cumulative sum of released mass up to some particular reference LS ratio (Eq. 2.2) is considered and expressed as cumulative concentration (Eq. 2.3). In Germany, e.g., the LS ratio of 2 is applied in "compliance tests", which are the everyday procedure to assure the quality of recycling materials (Kalbe et al., 2008; Susset & Grathwohl, 2011; Finkel & Grathwohl, 2017; Prieto-Espinoza et al., 2022). The cumulative mass released defined as:

$$M_{\text{released},LSX} = \sum_{i=1}^{n_{frac}(LS=X)} C_{frac,i} \cdot V_{w,i}, \tag{2.2}$$

where $M_{\text{released},LSX}$ is the mass released until a specific LS value X is achieved, $C_{frac,i}$ represent the measured actual concentration in a certain LS fraction and $V_{w,i}$ indicated the water volume of the individual fraction, and $n_{frac}(LS = X)$ is the number of fractions sampled until $LS = X$. Furthermore, the cumulative concentration at LSX, C_{LSX}, is defined by:

$$C_{LSX} = \frac{M_{\text{released},LSX}}{\sum_{i=1}^{n_{frac}(LS=X)} V_{w,i}}. \tag{2.3}$$

2.2.2 *Compilation of Available Data*

Over the last two decades, data from 82 extended column tests with CDW were compiled in accordance with the German standard (DIN 19528, 2009) at various locations in Germany. The data set used in this study is presented in Susset, Leuchs, 2008a. Measurements were taken at LS values of 0.3, 1, 2, and 4 in all 82 tests. Data at LS=10 is only available from 23 tests. Each of the 351 measurement events includes 9 parameters: Sulfate, Vanadium, Chromium, Copper, 15 (US-EPA) PAHs, as well as Dissolved Organic Carbon (DOC), pH, Electrical Conductivity (EC), and the LS ratio. The data set comprises a total of 3159 individual data points.

The distribution of the relevant chemical compounds' concentration in terms of kernel probability density for different ranges of LS is shown in Figure 2.1. See Figure B.1 for the distribution of pH, EC, and DOC in the data set.

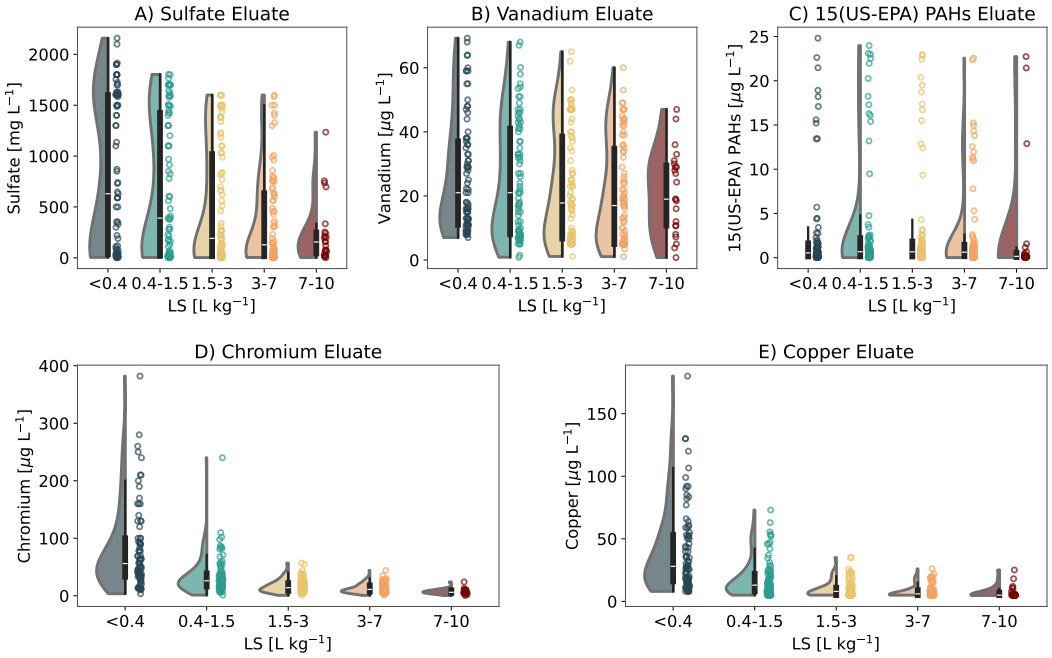


Figure 2.1: Half-violin plots, embedded with data points, represent kernel density estimates of the full distributions of the concentration in the leachate (i.e., eluate) of CDW in relevant LS ranges: A) Sulfate, B) Vanadium, C) 15 (US-EPA) PAHs, D) Chromium, E) Copper; note that the box inside the violin shapes represents the minimum, first quartile, median (white dot inside the box), third quartile, and maximum data value.

2.2.3 *Data Preparation*

Careful data preparation in terms of preprocessing and cleaning the data for subsequent detailed processing is crucial for the performance of ML algorithms. It essentially includes proper consideration of missing data values, feature scaling, and splitting the data into training and test sets (Dorj & Altangerel, 2013; Sarker, 2021).

Missing Data

At high LS, the dilution of highly soluble substances with limited mass in the solids may lead to concentrations below the quantification limits (Susset & Grathwohl, 2011). It is distinguished between true missing values (measurement has not been done), e.g., unmeasured 15 (US-EPA) PAHs values at low LS ranges (53 data points), and values below the Limit Of Quantification (LOQ) (i.e., predefined measurement device's detection limit) which included Copper (68 data points), Chromium (20 data points), and Vanadium (14 data points). For handling the values below LOQ, the approach of Faucheux et al., 2021 has been followed, replacing missing data with the LOQ of the respective chemical compound (i.e., the LOQ values for Copper, Chromium, and Vanadium were 5 µg/l). Moreover, due to the high accuracy and simple implementation, the polynomial extrapolation method has been used to fill in unmeasured 15 (US-EPA) PAHs values (Lepot et al., 2017).

Feature Scaling

ML algorithms' attempt to find trends in the data may be impeded, if data feature values (concentrations of Sulfate, Vanadium, Chromium, Copper, 15 (US-EPA) PAHs, DOC, and pH, EC, LS) have considerably different scales. Re-scaling of feature values into a fixed range is a common way to avoid this. In this work, whole data is re-scaled into the range [0,1] using the min-max normalization method (see Eq. A.6).

Training and Test Set

In developing a robust ML model, the data is split into two parts, training and a test set (Xu & Goodacre, 2018), typically with an 80/20 split, both sets being produced by uniformly random sampling from the preprocessed data (Lever et al., 2016). The training set is used by the ML algorithm to learn general patterns and features in the data set. The test set represents "unseen" data, which is used in the testing process to evaluate the performance of the previously trained model with its optimized hyperparameters obtained in the training process.

2.2.4 *Machine Learning Models*

Several algorithms (*Ordinary Least Squares*, *LARS-LASSO*, *Decision Tree*, *ANN*, *LASSO*, *Ridge*, *Extremely randomized Trees (ET)*, and *Random Forest (RF)*) were initially explored to predict the concentrations of the five relevant compounds at LS=2, 4, and 10, as well as the cumulative concentration at LS=2. After thorough testing, we narrowed down our selection to four ML models based on their R-squared and Cross-Validation performances. The two linear models, LASSO and Ridge, showed better predictive abilities compared to the other linear models we tested, and were therefore included in our final selection of models. Among the non-parametric models, we chose RF and ET due to their good performance in preliminary testing.

A maximum of 18 data features have been used as the input of each model measured at early testing times, i.e., LS=0.3 and LS=1, as depicted in the Data Features of Figure 2.2. Due to the limited availability of measurement data at LS=10 in the database, the main study relied on models capable of concurrently estimating two output variables - specifically, the concentrations of the respective compounds at LS values of 2 and 4. The procedure that was followed to develop, train, and evaluate these models is illustrated in Figure 2.2. Section B.8 provides details about the coding environment and the computer system utilized for developing the models.

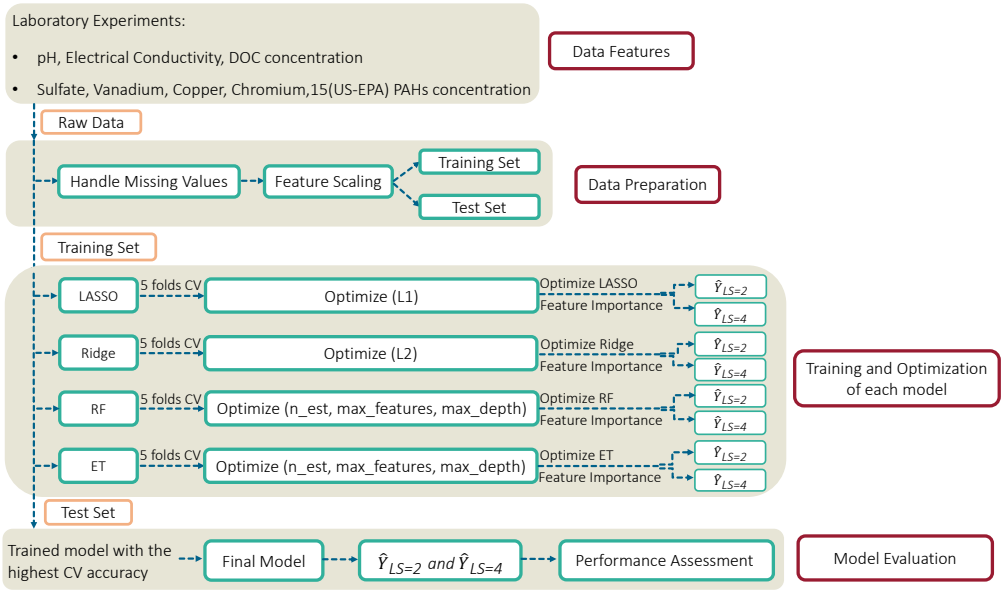


Figure 2.2: Flow chart of the development and assessment of compound-specific multi-output ML models.

Furthermore, to predict leachate concentrations at LS=10, three approaches to develop/-train ML models were compared: (i) "Early-stage input" models using data observed at LS=0.3 and 1 as model's input - like the models used to predict the concentration at LS=2 and 4, (ii) "Sequence-timepoint input" models using an extended input from observed data at LS=0.3, 1, 2, and 4, and (iii) "Hybrid input" models with input both from observed (at LS=0.3 and 1) and predicted data (at LS=2, and 4).

2.2.5 *Regression Analysis*

In ML, both classification and regression are common types of problems that are used to develop models. A regression model is a type of model that is used to predict a continuous outcome using one or more hyperparameters to control model complexity. Hyperparameters are the parameters of a ML model that are not learned from the data during training. They control the complexity and behavior of the model, and selecting the optimal hyperparameters is essential for achieving good performance. For this study, we have selected common linear and ensemble regression methods with several hyperparameters to compare the performance of the two different types of regression algorithms, which are presented shortly in the following:

Linear Regression

Linear regression is a simple yet powerful model for predicting a target variable, but it is prone to overfitting in high-dimensional settings. Regularization techniques like Ridge and Lasso help mitigate this issue and improve generalization (Pedregosa et al., 2012). *LASSO* employs the (L1) regularization method to eliminate less essential features by penalizing the "Absolute Value of the Magnitude of Coefficients" to induce the specific coefficients to be absolute zero (Tibshirani, 1996). *Ridge* regression employs (L2) regularization, defined

as "Squared Magnitude of Coefficients", which leads to minor weights and never sets the coefficient value to absolute zero (see details in A.6).

Ensemble Regression Methods

To overcome the limited generalizability of a single estimator model like linear regression, various ensemble methods have been developed that combine the predictions of multiple estimator models (Dietterich, 2000; Schapire & Freund, 2012). Two popular averaging ensemble methods are Random Forest and Extremely randomized Trees, which have been widely used in ML applications. The following briefly explains these methods:

Random Forest (RF) regression is a meta-estimator using several numbers of trees as an estimator model (n-est) with a certain depth (max-depth) using sampling with replacement (e.g., a bootstrap sampling— Section B.9) and a random subset of features (max-features), which leads to a reduction of the variance and overfitting of a model, yielding an overall better model performance (Breiman, 2001).

Extremely randomized Trees (ET) employ the same hyperparameters (n-est, max-depth, and max-features) as RF. However, to reduce the model's variance compared to RF, the randomness is taken one step further in how the features are split (Geurts et al., 2006). Instead of looking for the most discriminative thresholds, arbitrary thresholds are chosen for each candidate feature (Hbali et al., 2018). The best of these randomly generated thresholds is chosen as the splitting node. This difference can lead to ET having a higher model bias than RF. Although this approach may not capture the most critical relationships in the data as effectively as RF, it can make ET less prone to overfitting and better able to handle complex data relationships (Geurts et al., 2006). Refer to A.7 for more details.

2.2.6 *Hyperparameters Optimization: Cross-Validation*

In this study, models' hyperparameters were optimized using the random search method (for detailed information, refer to Section B.10), which involves training the model multiple times with random combinations of hyperparameters and evaluating its performance using the Cross-Validation (CV) technique (Bergstra & Bengio, 2012). The larger the number of trained model hyperparameters is, the better is typically the performance of the trained model (Xu & Goodacre, 2018). However, higher model complexity generally increases the risk of overfitting, which prevents the model's generality and leads to a rather poor model performance for unseen data (Ying, 2019).

To overcome the issue of overfitting, each model is trained on the CV set using the selected range of hyperparameters (as described in Table B.2) to optimize the model's complexity through the tuning of hyperparameters, resulting in a more stable and reliable model (Refaeilzadeh et al., 2016). In order to develop a CV set, the training data are split into K folds, where each fold is used as a validation set once, and the model is trained on the remaining K-1 folds. This process is called K-fold Cross-Validation.

Repeated K-fold CV was utilized in this study as a variant of K-fold CV, whereby the K-fold procedure was repeated a specified number of times (n), and the folds were split differently in each repetition. This approach is known to enhance the stability of results and reduce the variance of the performance estimate, particularly in limited data situations (Berrar, 2019). The tuned parameters of each algorithm were selected based on their highest averaged predictive performance (Figure 2.3).

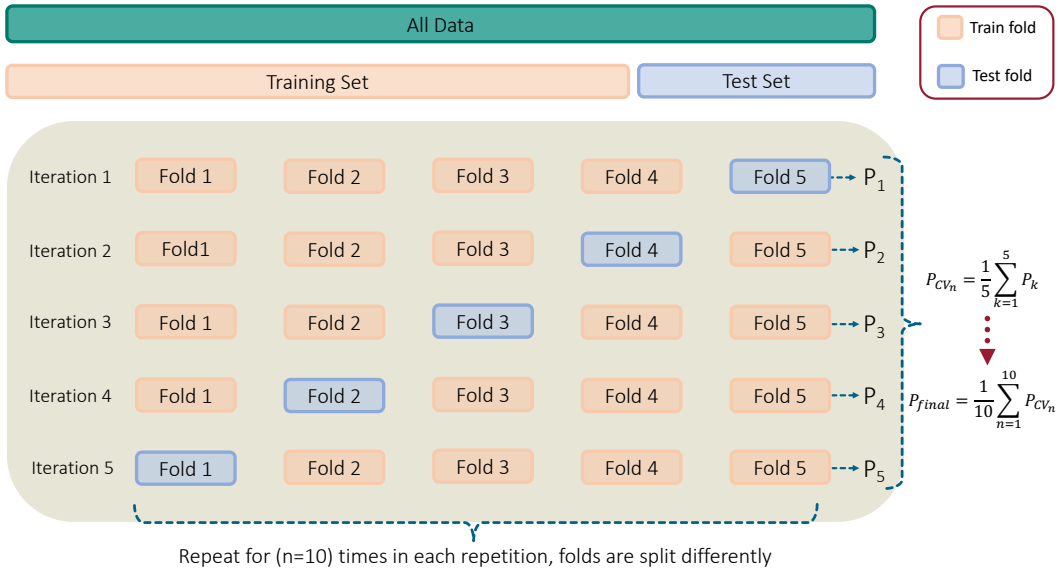


Figure 2.3: Repeated Five-Fold Cross-Validation (CV) chart. In each iteration, the training set is split into five parts (K=5), four of them (orange boxes) are taken as training data, and one fold is utilized as test data for evaluation (blue boxes). This CV procedure is repeated for (n=10) times, which n is the number of repetitions of the K-Fold procedure. The average value of the five iterations' test results counts as a performance of a specific K-Fold CV, and the average of ten repeated K-Fold CVs is computed as an assessment of the final repeated K-fold CV set.

2.2.7 *Feature Importance*

To quantify the importance of individual data features used as input to the ML models, we employed the *SHapley Additive exPlanation (SHAP)* algorithm (Lipovetsky & Conklin, 2001; Hazra & Anjaria, 2022). The algorithm aims to identify the contribution of each feature in the data set to the final prediction made by the model (Shapley, 1953). This information can be used to improve the model's performance and interpretability.

The SHAP algorithm is based on the concept of Shapley values from game theory (Shapley, 1953). In the context of ML, each feature in the data set is considered a player in a cooperative game, and the algorithm calculates the contribution of each feature to the final prediction by averaging its contribution across all possible combinations of the input features. This approach provides a fair attribution of each feature's contribution to the prediction, even for models with highly correlated input features (Lundberg & Lee, 2017). The SHAP value is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)], \tag{2.4}$$

where F is the entire set of model input features, S is a subset of F, $|S|$ and $|F|$ are the size of S and F, respectively. f_S is the prediction made by the model with the features in S. Also, $f_{S \cup \{i\}}$ refers to the prediction of the model when given a subset of input features (S) and a single input feature (i). The sum is taken over all possible subsets S of F that do not include i, and x_S describes the values of the input features in the set S. The term $|S|!(|F|\text{-}|S|\text{-}1)!$ is

a combinatorial factor that accounts for the number of ways that the features in S can be permuted.

2.2.8 *Model Performance*

For each of the relevant chemical compounds, the quality of the predictions made by the four different models (based on LASSO, Ridge, RF, and ET algorithm) was evaluated for the training and test data set as well as in terms of average performance in the CV set (Figure 2.3).

Two standard metrics, Normalized Root Mean Squared Error (NRMSE) and Coefficient of determination (R^2), were used in order to evaluate the performance of individual ML models. NRMSE and R-squared were calculated for each compound-specific model and predicted concentrations at LS=2, 4, and LS=10 separately as:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}}{\overline{y}} \tag{2.5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{2.6}$$

where $\hat{y}_i$ represents the predicted value for the ith observation, y_i depicts the measured value for the ith observation, $\overline{y}$ defines the average of observation, and n describes the number of observations considered in the evaluation. To predict the concentration at LS=2 and LS=4, a training set of n=67 data points is used. However, for predicting the concentration at LS=10, a smaller training set of only n=16 data points is utilized. Once the model is trained, it is tested on a separate test set to evaluate its performance. This test set consists of n=15 data points for predicting the concentration at LS=2, 4, and n=6 data points for predicting the concentration at LS=10.

Finally, the accuracy of the models was evaluated by separately calculating the mean NRMSE and R^2 values for the predicted concentrations at LS=2, 4. To assess the performance of the model at LS=10, the observed and predicted values at LS=10 were directly compared, and the corresponding NRMSE and R^2 values were calculated.

2.3 RESULTS AND DISCUSSION

2.3.1 *Overall Performance of the Models at LS=2 and 4*

All four tested algorithms perform reasonably well for each of the relevant compounds. The compound-specific performance of each optimized model (see Table B.2 for the tuned hyperparameters) in predicting the actual concentration at LS=2 and LS=4 on the test data set is given in Table 2.1.

	LASSO				Ridge				RF				ET			
	R^2_{train}	R^2_{CV}	R^2_{test}	$NRMSE_{test}$	R^2_{train}	R^2_{CV}	R^2_{test}	$NRMSE_{test}$	R^2_{train}	R^2_{CV}	R^2_{test}	$NRMSE_{test}$	R^2_{train}	R^2_{CV}	R^2_{test}	$NRMSE_{test}$
Sulfate	0.92	0.85	0.87	0.41	0.92	0.86	0.87	0.40	**0.98**	**0.92**	**0.94**	**0.27**	0.99	0.91	0.93	0.28
Vanadium	0.95	0.90	0.89	0.22	0.96	0.89	0.92	0.17	0.98	0.89	0.92	0.19	**0.99**	**0.92**	**0.97**	**0.12**
Chromium	0.84	0.70	0.23	0.29	0.86	0.70	0.09	0.32	0.95	0.68	0.75	0.17	**0.98**	**0.72**	**0.82**	**0.16**
Copper	0.83	0.73	0.89	0.17	0.86	0.68	0.48	0.39	0.93	0.72	0.90	0.17	**0.94**	**0.75**	**0.92**	**0.15**
15 (US-EPA) PAHs	0.95	0.89	0.94	0.42	0.95	0.88	0.94	0.4	**0.99**	**0.93**	**0.98**	**0.29**	0.99	0.92	0.97	0.32

Bold values represent the performance of the final model with the highest Cross-Validation R^2.

Table 2.1: Overall performance of each model to predict actual concentration at LS=2 and LS=4 using the best-tuned algorithms.

For each compound, the model with the highest R^2 score for the CV data set was selected as the *best available model*. All five models show high performance with R^2 scores > 80% and NRMSE < 0.3. The comparison between observed concentration values of the test data set and model predictions illustrates the good performance of these models (Figure 2.4).

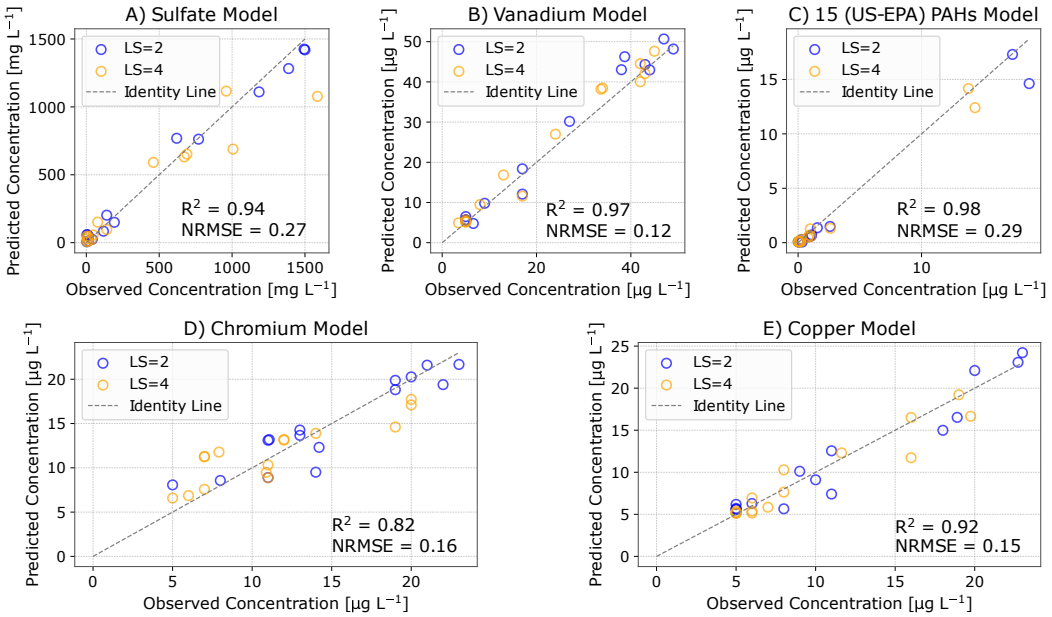


Figure 2.4: Measured vs. predicted concentrations at LS=2 and LS=4 for the test set samples using the best-available ML models for A) Sulfate, B) Vanadium, C) 15 (US-EPA) PAHs, D) Chromium, E) Copper (see Figure B.4 in B.7 for the performance of ML models in predicting the cumulative concentration at LS=2).

Predictions for Sulfate, Vanadium, and the 15 (US-EPA) PAHs are generally better than for Chromium and Copper. Two reasons were suspected for the comparatively moderate performance of the Chromium and Copper models in the CV set: data bias due to the substitution of below-LOQ values and variability in leaching behavior. For both Copper and Chromium, a notable number of data points (68 and 20, respectively) were set by substitution of originally below-LOQ values in the high LS range (e.g., LS=4) of the training data set. Obviously, such a substitution introduces some bias, in the sense that the substituted data does not perfectly represent the real leaching process. We compared three options for substitution (by (i) a fixed value of LOQ, (ii) a fixed value of LOQ/2, and (iii) random values between 0 and LOQ) but did not see any difference in the performance of the models.

Thus, data bias due to substitution does not seem to significantly affect the learning of the models.

The variability in the leaching behavior of specific compounds was further examined by analyzing the performance of all individually trained models during the CV. The results showed considerable variability in the performance of Chromium and Copper models from one training-validation iteration to the next, most likely due to the variability in the data representing different leaching behavior in individual validation folds (see Figure B.3). Consequently, the model becomes too closely tied to the specific characteristics of the training data, resulting in poor generalization of unseen data.

These ML models also demonstrated high accuracy in predicting the cumulative concentrations at LS=2 ($C_{LS2,m+p}$) for all relevant compounds (Figure B.4). The cumulative concentrations were calculated using modified equations (2.2) and (2.3), which take into account both the measured (m) and predicted (p) concentrations:

$$C_{LS2,m+p} = \frac{(\sum_{i=1}^{n_{frac,m}} C_{frac,i(m)} \cdot V_{w,i}) + (\sum_{i=n_{frac,m+1}}^{n_{frac,m+p}} C_{frac,i(p)} \cdot V_{w,i})}{\sum_{i=1}^{i=n_{frac,m+p}} V_{w,i}}, \tag{2.7}$$

where $n_{frac,m}$ =2 (at LS=0.3 and LS=1) and $n_{frac,m+p}$ =3 (at LS=2). This approach is particularly important as German regulations categorize recycling materials in terms of their re-usability in infrastructural and landscaping measures, based on cumulative concentrations at LS=2.

2.3.2 *Overall Performance of the Models at LS=10*

To evaluate the performance of the models in predicting the concentration of relevant compounds at LS=10, three different approaches were used. The results, as shown in Figure 2.5 and summarized in Table B.3, indicate that the highest accuracy in predicting the concentration at LS=10 was achieved by the "Sequence-timepoint input" model (using observed concentration at LS=0.3, 1, 2, and 4 as input). The "Hybrid input" model (using observed concentration at LS=0.3 and 1, and the predicted concentration at LS=2 and 4 as input) had a slightly lower accuracy. It is remarkable that this model's performance for each specific compound was comparable in predicting the concentration at LS=10 to that of the "Sequence-timepoint input" model, despite using the predicted concentrations at LS=2 and 4. The lowest accuracy in predicting the concentration of relevant compounds at LS=10 was achieved by the "Early-stage input" model (using observed concentrations at LS=0.3 and 1 as input).

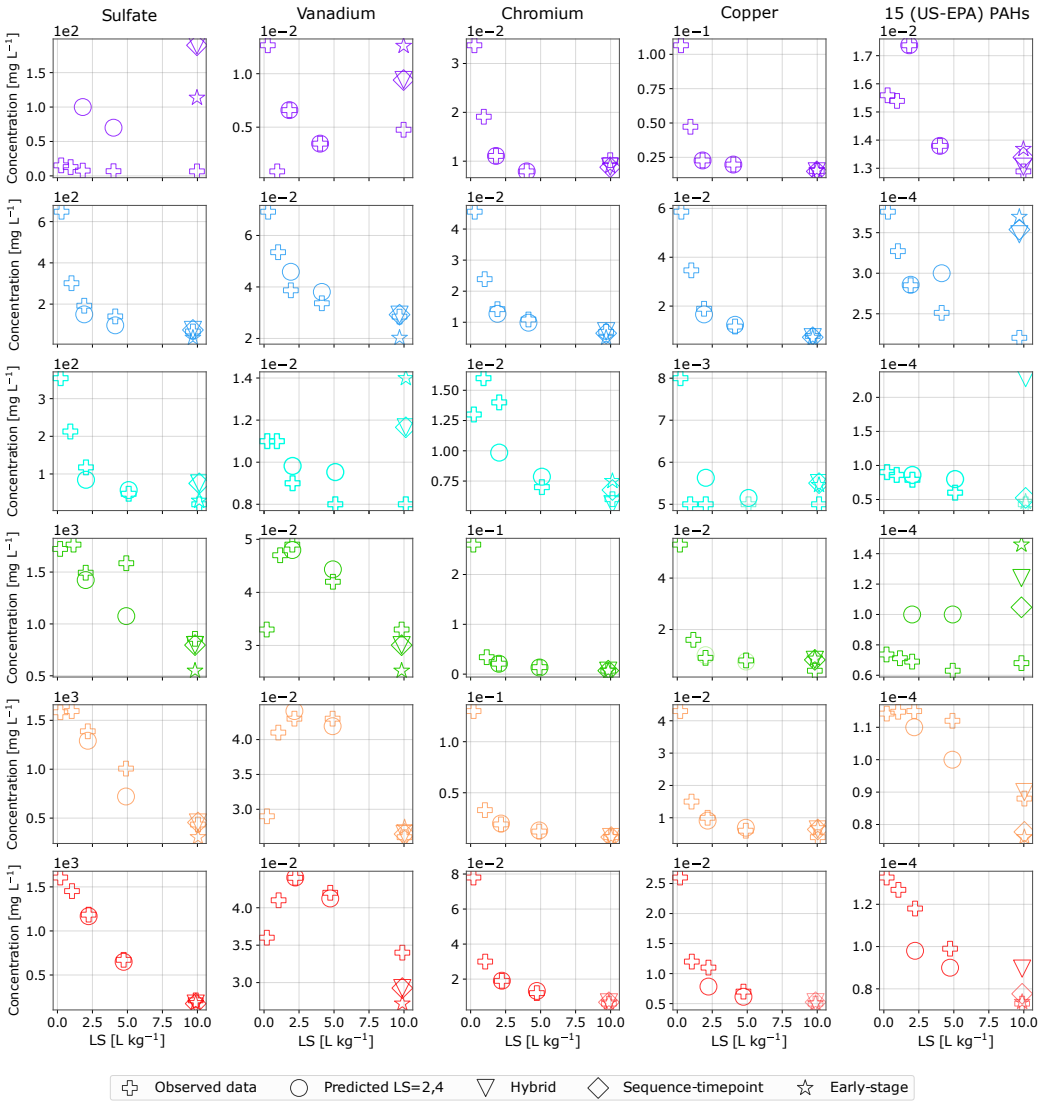


Figure 2.5: Leaching behavior of six different samples assessed by various ML models. Each column represents a specific compound, and each row with different color shows a specific sample. Measured data is depicted by the plus symbol, while predicted concentrations at LS=10 by sequence-timepoint, hybrid, and early-stage input models are represented by a diamond, reverse triangle, and star symbol, respectively. Predicted concentrations at LS=2 and 4 by the ML model (using LS=0.3 and 1 as input to predict concentrations at LS=2 and 4) are shown by circles.

These results suggest that incorporating data at higher LS values in the training data leads to improving the accuracy of the ML models in predicting the concentration of the relevant compound at LS=10. Moreover, the use of predicted values from the ML model at LS=2 and 4 as inputs to the second model improved the performance of the model compared to using only observed data at LS=0.3, and 1. While the performance for the training and test sets of these models was relatively high, it is important to note that overfitting may be a concern due to the limited number of available training samples. This is reflected in the very low R^2 score observed in the CV set for these models (Table B.3).

2.3.3 *Influence of the Size of the Training data set on Model Performance*

ML algorithms' success is attributed to the availability of extensive training sets. The performance of resulting ML models typically improves with the growing number of available samples in the training data set (Jha et al., 2019). This could be confirmed for the performance of the compound-specific models as a function of the training set size. As illustrated in Figure 2.6, the models' performance is low when the training data set size is small. With increasing training data set size, the models' performance improves, inferring that the models learn to find the general pattern from the training set. Results for Chromium indicate a larger variation in the leaching behavior as represented by each new test sample. The effect of training data set size is evident for these compounds. In contrast, the leaching behavior of Sulfate appears to be represented very well even by very limited training data (due to its relatively stable behavior typically showing a rapid decrease of leachate concentration). The Sulfate model thus shows very good performance if trained with a small data set.

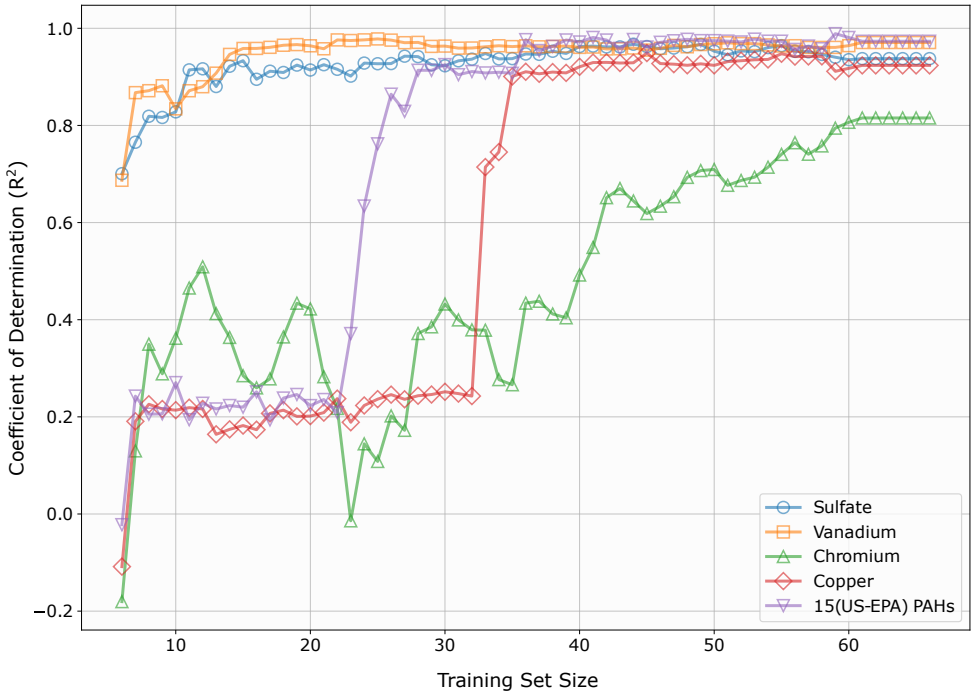


Figure 2.6: Influence of training data set size on the performance (R^2) of ML models to predict concentrations at LS=2 and 4. The best-available model for each compound is used. While maintaining the structure of the models (optimized hyperparameters), the models were re-trained for different sizes of the training data set and evaluated for their performance based on the test data set. Note: The maximum value of R^2 is 1 but has no minimum value; a negative R^2 means that the model is predicting worse than the mean of the target values.

2.3.4 *Key Model Features*

The importance of all 18 models' features (ML model to predict concentration at LS=2 and 4) was analyzed with the SHAP method. As could be expected, data from early measurements (respective compounds concentration at LS=0.3 and LS=1) do most significantly

affect the outcome (predicted concentrations at LS=2 and LS=4) of all five models (Figure 2.7). In addition to these primary key features, a number of secondary key features were identified for each model. These features can be reasonably linked to the release specifics of the respective compound. For example, as illustrated in Figure 2.7, the results indicate that pH and EC have a significant impact on predicting Sulfate concentration, which corresponds to these parameters' potential to strongly influence the aqueous solubility of Sulfate. Figure B.1 shows the pH distribution of all 82 samples, which lies within the range of 8-13. According to Engelsen et al., 2017, the leachability of Sulfate decreases at higher pH values above 8. Moreover, higher EC values may promote the dissociation of Sulfate into their constituent ions, resulting in increased Sulfate concentrations.

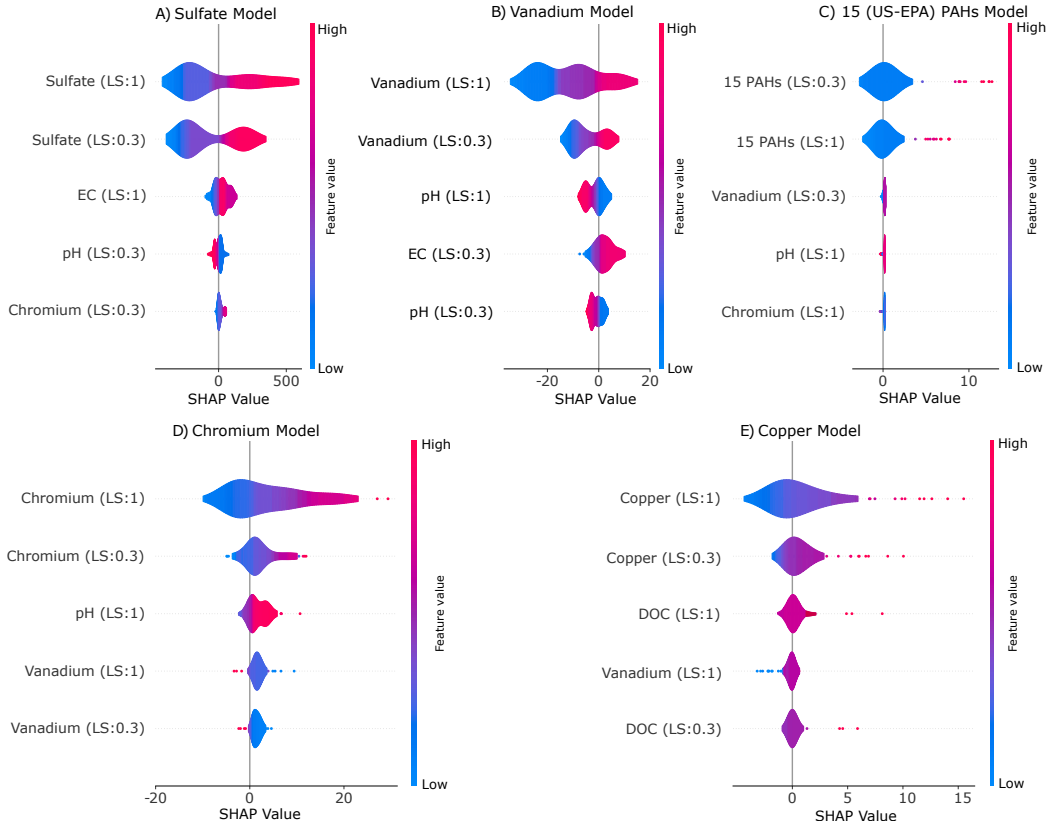


Figure 2.7: Beeswarm summary plots of SHapley Additive exPlanation (SHAP) values ϕ. Features are ranked according to their importance for the 5 most important features: the higher the feature is placed in the plot, the more important it is for the model outcome (in terms of mean absolute ϕ value). For each feature, a single dot refers to the calculated ϕ value using the data x_S of a single column leaching test (see Eq. 3.8). Feature importance varies in terms of magnitude and direction of correlation. Dependency on the feature's value is indicated by color.

Although pH and EC are recognized as key factors in determining the solubility of Sulfate, their impact on the leachability of Vanadium can be more intricate and may be influenced by the specific oxidation state of Vanadium in a given sample (Peng, 2019). Nonetheless, pH and EC are crucial determinants in regulating the leaching behavior of several metals and metalloids. Environmental risks related to the mobility of Vanadium typically decrease at higher pHs (Engelsen et al., 2012), as confirmed by the SHAP analysis shown in

Figure 2.7, whereas higher EC values can enhance the desorption and liberation of metals and metalloids. (Huang et al., 2015)

In addition, Del Rey et al., 2015 found that a considerable proportion of total Chromium released from CDW exists in the form of Cr (VI), which is known to be more harmful and easily leached than Cr (III) due to its pH-dependent oxidation state and leachability. It is important to note that the leachability of Chromium increases between pH levels 8-12 and leached as Cr (VI) (Engelsen et al., 2012; Del Rey et al., 2015). This is consistent with the findings of the SHAP analysis, which also investigated the behavior of Chromium (Figure 2.7).

As shown in Figure 2.7, DOC can have a significant impact on Copper leaching in demolition waste. This can be achieved by forming soluble complexes with Copper ions. This can increase the mobility and bioavailability of Copper in the environment (van Zomeren & Comans, 2004; Mesquita & Carranca, 2005). Factors like the concentration and type of organic matter, pH, and presence of other metal ions can affect the complexation of Copper by DOC (Meima & Comans, 1999).

2.4 SUMMARY AND CONCLUSION

In this study, we have demonstrated the suitability of using Machine Learning (ML) models to predict the long-term leaching of relevant compounds from Construction and Demolition waste (CDW). By utilizing short-term concentration data (LS<=1), the models were able to accurately predict long-term concentrations (LS=2, 4, and 10), resulting in a significant reduction in the time required for laboratory testing from nearly 7–14 days (depending on the leaching test standard) to just 1 day.

Time-consuming long-term measurements may therefore be replaced by predictions with ML models, which showed to be particularly feasible in the context of cumulative concentration levels used in Germany to categorize recycling materials with respect to their re-usability in infrastructural and landscaping measures (Bundesgesetzblatt, 2021). Values of predicted cumulative concentration at LS=2, agree very well with measured cumulative concentrations (Figure B.4). Furthermore, the replacement of long-term measurements by predictions does not compromise a correct categorization of CDW. For all of the column test data considered in the ML test set, the assessment of the CDW material in terms of its allocation into one of the material categories (CDW-1, CDW-2, CDW-3) would be appropriate if done based on the cumulative concentration. Specifically, it should be noted that all samples in the test set have Vanadium, Copper, and Chromium concentrations below the legal limits established by the German Recycling Decree (Bundesgesetzblatt, 2021) and fall into the CDW-1 class.

Finally, the limitations of this study have to be acknowledged. The amount of available data plays a crucial role in the performance of the ML models (see Figure 2.6). While the dataset comprising 82 column leaching tests was sufficient to reliably predict the long-term concentrations in CDW leachate, a dataset of similar size might not be enough for other materials, and a generalization of the findings may be difficult. More insight is expected from further studies, particularly from applications to soil materials, the properties and contaminant load of which are known to vary much more than for CDW. Also, in soil materials, there is a much larger variety of relevant compounds to be considered (a total of 18 compounds, incl. chlorinated compounds, and petroleum hydrocarbons; see Bundesgesetzblatt, 2021). Future research may also involve the consideration of meta-data as model

input. Data such as geographical provenance, time period of use, specifics of the recycling process, etc., may enhance model performance and interpretability of results.

The findings of this study have important practical implications for the management of waste and recycling materials, which are growing concerns in the construction industry. By providing a faster and more efficient method for evaluating the leaching behavior of relevant compounds, this approach could help to inform decision-making around the appropriate reuse or disposal of these materials.

3

ENSEMBLE SURROGATE MODELING OF ADVECTIVE-DISPERSIVE TRANSPORT WITH INTRAPARTICLE DIFFUSION MODEL FOR COLUMN LEACHING TEST

Amirhossein Ershadi[1], **Michael Finkel**[1], **Binlong Liu**[1], **Olaf.A Cirpka**[1], **Peter Grathwohl**[1]

[1]*Department of Geosciences, University of Tübingen, Schnarrenbergstraße 94-96, 72076 Tübingen, Germany*

Published in the **Journal of Contaminant Hydrology**
https://doi.org/10.1016/j.jconhyd.2024.104423

Author Contributions

First author:
Scientific ideas: 60%, Data processing: 90%,
Analysis & interpretations: 70%, Paper writing: 70%

3.1 INTRODUCTION

Column leaching tests offer crucial insights into the release and transport of contaminants in porous media (e.g., Grathwohl & Sloot, 2007; Kalbe et al., 2014). These tests assess the mass release of contaminants over a specific time period, expressed as the liquid-to-solid ratio (LS), that is, the volume of water (L) percolated through the column within a designated time period, in relation to the dry weight of the solids (S) contained in the column.

Process-based numerical-modeling approaches have been extensively utilized for simulating coupled advection, dispersion, and inter-phase mass transfer in column leaching experiments (Finkel & Grathwohl, 2017; Liu et al., 2021). Although these methods provide valuable insights into the leaching behavior of contaminants, they are not without certain drawbacks. Specifically, insufficient discretization of the domain can impede the accurate prediction of the concentration breakthrough curves. Furthermore, these methods often require a considerable allocation of computational resources, especially in sensitivity analysis and during model calibration, where parameters are adjusted to fit the model output to observed data.

In the context of model calibration, Bayesian methods offer a robust approach of estimating complete parameter distributions conditioned on observed data. Markov-Chain Monte Carlo (MCMC) methods, like the differential evolution adaptive metropolis algorithm DREAM of Vrugt, 2016 can estimate the full distributions of parameters conditioned on data, but require many calls of the predictive model for each data set. This is why machine learning techniques like Simulation-Based Inference (SBI) have been developed that perform the same estimation task at lower computational costs. SBI employs a subset of neural-network-based methods, particularly known as Neural Posterior Estimation (NPE). This approach involves training on model inputs and outputs that are sampled from the prior distribution, enabling the precise estimation of posterior parameter distributions when being confronted with data (Papamakarios & Murray, 2016; Lueckmann et al., 2017; Cranmer et al., 2020). This approach eliminates the need for additional inference procedures, naturally amortizing inference costs (Greenberg et al., 2019). Consequently, once the network is trained, inference for unseen measurements can be executed without requiring new simulations. Despite their effectiveness, these techniques encounter challenges, particularly the high computational costs associated with a large number of complex process-based model evaluations (often in the order of 10^4 simulations). To tackle this issue, a surrogate model may be employed to substitute the high-fidelity simulation model in the training of simulation-based inference.

Surrogate models, also known as emulators or proxy models, are powerful methods extensively utilized across various scientific and engineering domains, including subsurface hydrology. Among others, they have been used in parameter estimation for subsurface flow (Babaei et al., 2015; Allgeier & Cirpka, 2023), sensitivity and uncertainty analysis (Chen et al., 2021; Mohammadi et al., 2023), and as ensemble surrogate models in the domain of groundwater remediation (Jiang et al., 2015; Chu & Lu, 2015; Ouyang et al., 2017; Shams et al., 2021). Surrogate models approximate the relationship between model parameters and the output of complex models at low computational costs, typically employing Machine Learning (ML) methods, allowing the model behavior to be investigated and analyzed in a computationally efficient manner. Conversely, they may come at the cost of reduced accuracy in analysis.

The effectiveness of the surrogate model significantly relies on the quality and representativeness of the sample data utilized during its training phase, including the number and choice of initial samples, also denoted design of experiments (DoE) (Wang & Shan, 2007; Williams & Cremaschi, 2021). A good design of experiments involves generating a representative sample set that captures the essential features of the input-parameter space. This set of samples is utilized to train the surrogate model and guide the optimization process. By incorporating prior knowledge of the parameters and employing appropriate sampling methods, such as the space-filling Halton Sequence (Halton, 1960), an effective design can be constructed.

To enhance the accuracy and utility of individual surrogate models, optimization strategies are crucial, and Adaptive-Recursive-Sampling (ARS) represents a promising approach in this regard (Zou et al., 2007; Razavi et al., 2012). Adaptive sampling utilizes different sample selection criteria, including maximizing the Expected Improvement (EI) (Jones et al., 1998), maximizing the standard deviation (Liang et al., 2023), or employing distance-based metrics (Regis & Shoemaker, 2007, 2009; Allgeier & Cirpka, 2023). These methods intelligently adapt the sampling process based on the evolving behavior of the surrogate model, striking a balance between exploring areas of high uncertainty to prevent the algorithm from getting trapped in local minima and exploiting trust regions where the surrogate model is expected to show the optimal performance, minimizing the loss or objective function. By iteratively refining the surrogate model with targeted sampling, adaptive sampling reduces the number of evaluations required to achieve high accuracy by focusing on the most informative regions of the input space. This approach significantly enhances the performance of the surrogate model, enabling it to better capture the intricate behavior of the underlying process while lowering computational costs.

Although the accuracy of an independent surrogate model may be generally satisfactory within a specific range of parameters, creating an ensemble of surrogate models becomes essential for ensuring precise representations throughout the entire predefined parameter space. An ensemble surrogate model refers to combining a series of individual surrogate models to enhance overall predictive performance. Each base surrogate model is rooted in a distinct virtual-reality dataset (VRD) drawn from different subdomains of the prior parameter space (see Figure 3.1 for a simplified visualization in a 2-D parameter space). These virtual-reality datasets comprise simulated data points, each representing a specific parameter set. There are different techniques for constructing the ensemble surrogate model from the set of base surrogate models. The simplest approach consists of interpolation between different base surrogate models, e.g., by inverse-distance weighted interpolation or by Gaussian Process Regression (Su et al., 2017). In the stacking method (Wolpert, 1992), the base surrogate models are used to train yet another surrogate model, here using random forest as the ML technique. The former is conceptually simple and straightforward, whereas the latter allows the construction of a complex and potentially accurate ensemble model at the cost of intransparency in comparison to linear interpolation.

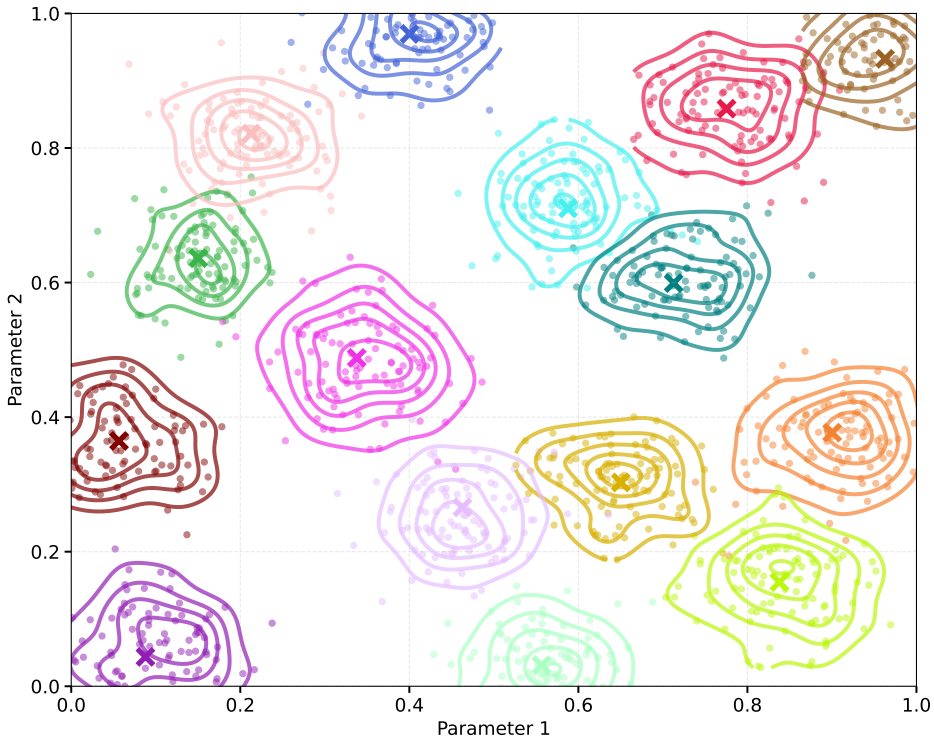


Figure 3.1: Exemplary representation of spatial distribution of the Virtual-Reality Datasets (VRD) (crosses) and derived training sets for base surrogate models (semi-transparent dots). Contour lines depict the frequency of data points, illustrating the distribution density of VRDs training set samples throughout the parameter space.

The primary objective of the present study is to develop an ensemble surrogate model that accurately mimics the physics-based simulation of column leaching tests throughout the space of parameters in the range of values observed in previous column leaching studies across many contaminants in many types of soils. These ensemble surrogate models emulate cumulative concentrations resulting from simulating advective-dispersive transport with intraparticle diffusion (ADE-IPD) in soil columns (Liu et al., 2021). We place specific emphasis on enhancing the simulation speed while obtaining practically the same results as the original model. By this, we want to facilitate the simulation of column leaching tests for soil contaminants in accordance with established standards set in Germany (DIN 19528, 2009). To optimize the individual surrogate model and employ an effective sampling approach, we utilized ARS, employing a combination of three different infill criteria. These criteria include maximizing the EI, maximizing the standard deviation, and staying below a threshold of the Mahalanobis distance to the currently best parameter set. To evaluate the impact of the initial sample quantity on the performance of each base model within the ensemble, results for two distinct DoE are compared. Finally, the proposed ensemble surrogate model serves as a cornerstone in the framework of model calibration for utilizing experimental data from column leaching tests to estimate the complete posterior parameter distribution. This is achieved through NPE, which provides full parameter distributions conditioned on data.

3.2 THEORY AND BACKGROUND

3.2.1 *Process-Based Column Leaching Model*

The German standard for column leaching tests (DIN 19528, 2009) defines two test steps (Figure 3.2): (1) saturation, in which clean water is injected into the dry contaminated soil until the lab column gets fully saturated, and (2) percolation, where water passes through the column and concentrations are measured in the effluent at different times.

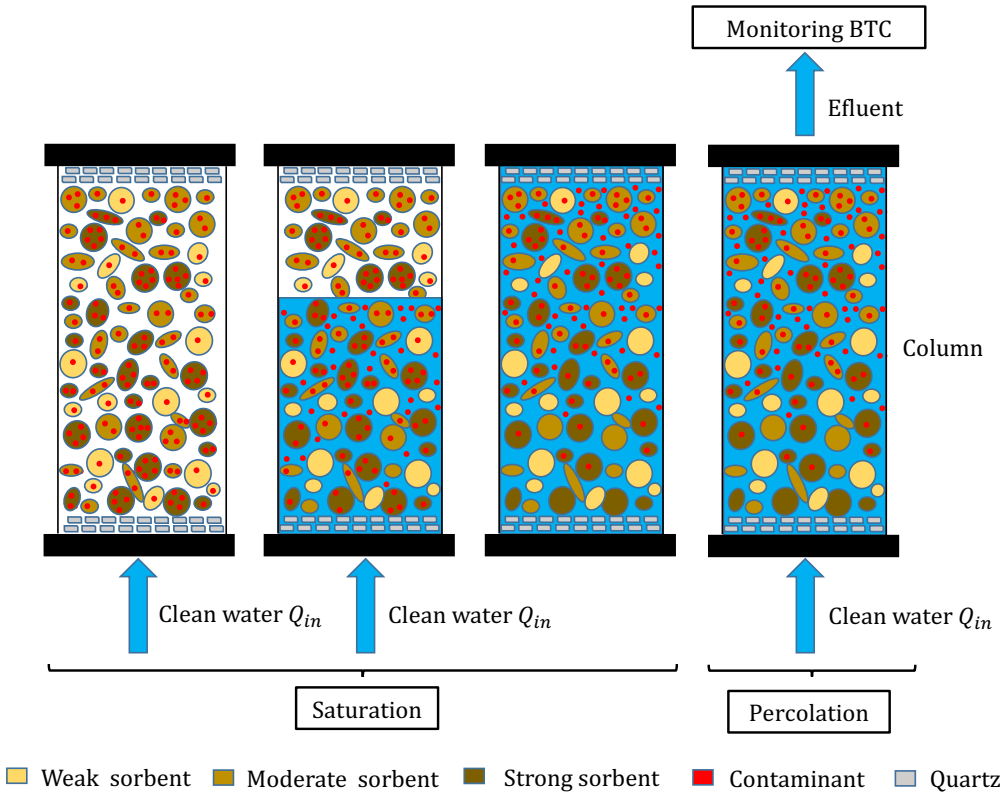


Figure 3.2: Flow chart of column leaching test of heterogeneous materials based on the German Standard (DIN 19528, 2009). Contaminants desorb from different sorbents and are transported by water, accumulating at the upper section of the column depending on mass transfer kinetics. The concentration of contaminants in the effluent is monitored over time, and breakthrough curves (BTC) are obtained.

To quantitatively describe column leaching tests, we assume (1) one-dimensional flow and advective-dispersive transport in the mobile pore space between the grains, (2) spherical porous grains of solids containing immobile water, (3) radial intraparticle diffusion in the intraparticle pore space, (4) local equilibrium of the concentrations between the intraparticle pore water and the sorbents following a linear isotherm. These assumptions lead to the 1-D advection-dispersion equation with intraparticle diffusion (ADE-IPD) (e.g., Liu et al., 2021):

$$\frac{\partial C_\mathrm{w}}{\partial t} = -v\frac{\partial C_\mathrm{w}}{\partial x} + D_\mathrm{L}\frac{\partial^2 C_\mathrm{w}}{\partial x^2} - \sum_{k=1}^{m}\left(\frac{3f_k(\varepsilon_k + \rho_{\mathrm{p},k}K_{\mathrm{d},k})(1-n)}{a_k^3 n}\right)\int_0^{a_k} r^2\frac{\partial C_{\mathrm{w,p},k}(r)}{\partial t}dr \quad (3.1)$$

$$\frac{\partial C_{\mathrm{w,p},k}}{\partial t} = D_{a,k}\left(\frac{\partial^2 C_{\mathrm{w,p},k}}{\partial r^2} + \frac{2}{r}\frac{\partial C_{\mathrm{w,p},k}}{\partial r}\right) \quad (3.2)$$

with the following boundary and initial conditions:

$$C_{\mathrm{w,p},k}(r = a_k) = C_\mathrm{w} \quad \forall t \quad (3.3)$$

$$\left.\frac{\partial C_{\mathrm{w,p},k}}{\partial r}\right|_{r=0} = 0 \quad \forall t \quad (3.4)$$

$$-D_\mathrm{L}\left.\frac{\partial C_\mathrm{w}}{\partial x}\right|_{x=L_\mathrm{col}} = 0 \quad \forall t \quad (3.5)$$

$$C_\mathrm{w}(x = 0) - \frac{D_\mathrm{L}}{v}\left.\frac{\partial C_\mathrm{w}}{\partial x}\right|_{x=0} = C_\mathrm{w,in} = 0 \quad \forall t \quad (3.6)$$

$$C_\mathrm{w}(t = 0) = 0 \quad \forall x \quad (3.7)$$

$$C_{\mathrm{w,p},k}(t = 0) = C_{\mathrm{w,p},k}(0) \quad \forall r \quad (3.8)$$

where $C_\mathrm{w}(x,t)$ [M L^{-3}] and $C_{\mathrm{w,p},k}(x,t,r)$ denote the aqueous-phase concentrations in the mobile water and in the intraparticle pore space of sorbent k, respectively; x [L], t [T], and r [L] are the spatial coordinate along the column axis, time, and the radial coordinate within the spheres, respectively. m [−] denotes the total number of sorbents. f_k [−], a_k [L], ε_k [−], $\rho_{\mathrm{p},k}$ [M L^{-3}], and $K_{\mathrm{d},k}$ [L^3 M^{-1}] are the volume fraction, radius, intraparticle porosity, dry bulk density, and distribution coefficient of sorbent k, respectively. v [L T^{-1}], n [−], and $D_\mathrm{L} = \alpha v + D_\mathrm{p}$ [L^2 T^{-1}] denote the seepage velocity of the water, the intergranular porosity, and the longitudinal dispersion coefficient. α [L], $D_\mathrm{p} = nD_\mathrm{aq}$ [L^2 T^{-1}], and D_aq [L^2 T^{-1}] denote the dispersivity, the pore diffusion coefficient, and the aqueous diffusion coefficient of the contaminant. $D_{\mathrm{a},k} = \frac{D_{\mathrm{e},k}}{\rho_{\mathrm{p},k}K_{\mathrm{d,b},k}} = \frac{D_\mathrm{aq}\varepsilon_k}{\tau_{\mathrm{f},k}\rho_{\mathrm{p},k}K_{\mathrm{d,b},k}} \approx \frac{D_\mathrm{aq}\varepsilon_k^2}{\rho_{\mathrm{p},k}K_{\mathrm{d,b},k}}$ [L^2 T^{-1}] is the apparent intraparticle diffusion coefficient with $K_{\mathrm{d,b},k} = K_{\mathrm{d},k} + \frac{\varepsilon_k}{\rho_{\mathrm{p},k}}$ being the bulk distribution coefficient between bulk water and porous particles, which also considers the solute mass stored in the intraparticle pore space; $\tau_{\mathrm{f},k}$ [−] is the tortuosity; $D_{\mathrm{e},k}$ [L^2 T^{-1}] is the effective diffusion coefficient. Empirical studies showed that $D_{\mathrm{e},k}$ increases approximately with the square of the intraparticle porosity (Boving & Grathwohl, 2001), so that the tortuosity can be approximated via the reciprocal of the intraparticle porosity, $\tau_{\mathrm{f},k} \approx \frac{1}{\varepsilon_k}$. L_col [L] denotes the column length; $C_\mathrm{w,in}$ [M L^{-3}] represents the contaminant concentration in the inflow of the column.

We solve the advection-dispersion equation with intraparticle diffusion using the cell-centered Finite Volume Method (Figure 3.3), in which the spherical particles are discretized into several spherical shells of equal volume (based on the method of Jaeger & Liedl, 2000). The column is spatially discretized into a number of cells with regular grid spacing, and the governing equations, Eqs. (3.1-3.8), are solved in time by implicit Euler integration using the Newton-Raphson scheme for linearization (for details see Liu et al., 2021).

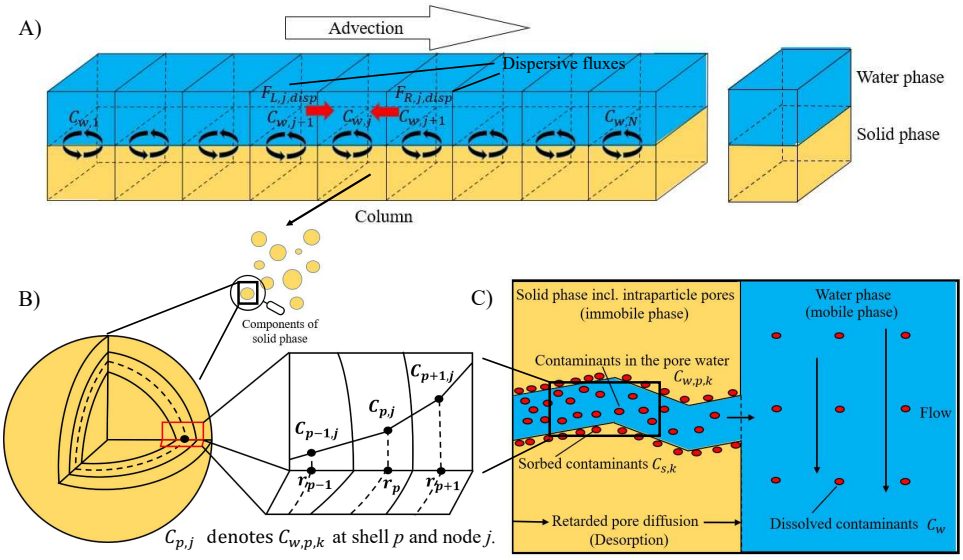


Figure 3.3: Numerical model setups for solving 1-D transport equation with contaminant desorption limited by intraparticle diffusion. (A) Discretization of the column into N cells. (B) Representation of the solid phase as a composition of grains having different sizes and properties, where each grain is discretized into certain shells. (C) Scheme of mass transfer limited by intraparticle diffusion in spheres (Liu et al., 2021).

Following German standard for column leaching tests (DIN 19528, 2009), we consider the cumulative concentrations $y(LS)$ [M M^{-1}] for given liquid-to-solid ratios LS [L^3 M^{-1}]:

$$LS(t) = \int_0^t \frac{Q(\tau)}{m_s} d\tau \tag{3.9}$$

$$y(t) = \int_0^t C_{w,\text{out}}(\tau) d\tau, \tag{3.10}$$

where $Q(t)$ [L^3 T^{-1}] is the flow rate at time t.

3.2.2 *Surrogate Models*

Estimating the parameters of a model from data requires the definition of a loss or objective function, such as the sum of squared differences between model outcomes and measurements. In many real-world problems, the evaluation of the objective function is computationally expensive due to complex simulations or extensive data processing. Surrogate modeling, combined with ARS, offers a powerful approach to address such challenges (Razavi et al., 2012). In this study, the terms adaptive-recursive-sampling and adaptive-sampling are used interchangeably. Adaptive sampling markedly improves the efficiency of the objective function evaluation through a recursive sampling strategy within the region of interest. This involves iteratively sampling new points, evaluating them using the objective function, and subsequently updating the surrogate model. In this study, the "objective function" is expressed in the form of the relative root mean squared error (RRMSE):

$$f_{\text{obj}}(\boldsymbol{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{\hat{y}_i}{y_i}\right)^2}, \tag{3.11}$$

where f_{obj}, y and $\hat{y}$ represent the objective function, the actual and surrogate model's predicted cumulative concentrations based on the input parameter set $\boldsymbol{\theta}$, respectively, with the actual concentrations (i.e., physics-based model output) derived from the VRD. The variable n denotes the number of measurement points, which, according to the German standard (DIN 19528, 2009), comprises cumulative concentrations at 7 different liquid-to-solid ratios: 0.1, 0.2, 0.5, 1, 2, 5, 10 [L kg^{-1}].

By dynamically improving the surrogate model through recursive sampling, adaptive sampling optimally navigates the search space, providing a powerful method for enhancing the overall efficiency and accuracy of the objective function evaluation process. The flowchart of Figure 3.4 illustrates the general process of surrogate modeling.

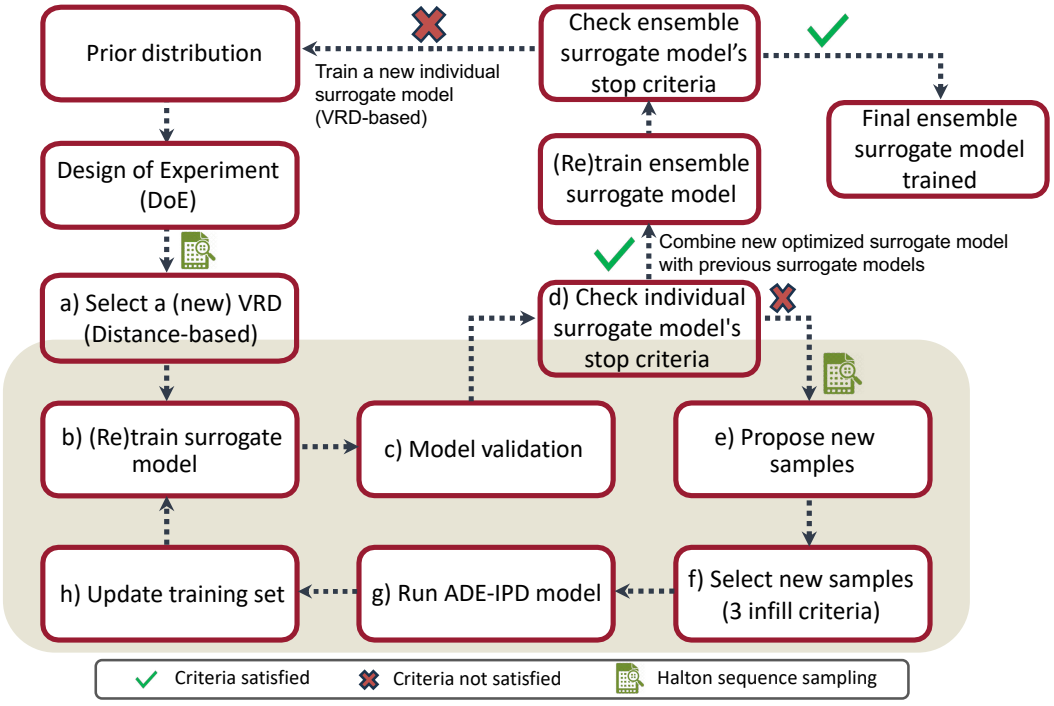


Figure 3.4: Flowchart illustrating the surrogate modeling process for advective-dispersive transport with intraparticle diffusion (ADE-IPD) in a column leaching model. The beige box represents the adaptive-recursive-sampling loop for an individual surrogate model, which is based on a specific Virtual-Reality Dataset (VRD).

1. The "Prior Distribution" expresses our initial understanding of the system, here represented by an uncorrelated uniform distribution of each parameter within a range that is in accordance with literature values (see Table 3.1).

2. The "Design of Experiments (DoE)" method is strategically employed to carefully arrange data collection by generating an initial set of samples (i.e., training set) within the bounds of the prior distribution. This is achieved by utilizing a specific sampling technique, in our case, the Halton Sequence (see section 3.2.2).

3. Adaptive-Recursive-Sampling:
    a) "Select a New Virtual-Reality Dataset (VRD)" from a constant pool of 200 unique VRDs, which were generated only once using Halton Sequence sampling, to serve as the data foundation for training an individual surrogate model. This selection is guided by either the maximum or minimum Mahalanobis distance metrics (see section 3.2.2).
    b) "(Re)training the Surrogate Model" involves the development or improvement of a machine learning model by integrating the initial and newly acquired data from adaptive sampling to enhance the surrogate model's accuracy for a specific VRD. In this study, we used the Extremely randomized Trees (Extra-Trees) algorithm to create individual surrogate models for each VRD within the adaptive-sampling loop (see section 3.2.2).
    c) "Model Validation" constitutes of a rigorous evaluation of the individual surrogate model's quality and precision. This assessment centers on scrutinizing the RRMSE, specifically in relation to the VRDs of cumulative concentration for each liquid-to-solids ratio within every surrogate model (see section 3.2.2).
    d) "Check Stopping Criteria" involves the evaluation of predefined criteria to ensure efficient resource utilization and goal attainment when ending the adaptive sampling loop (see section 3.2.2).
    e) "Propose" 5,000 new samples using Halton Sequence sampling to extend the previously generated proposed samples (see section 3.2.2).
    f) "Select" 100 samples according to specific infill criteria for exploitation and exploration (see section 3.2.2).
    g) "Run ADE-IPD Model" (i.e., the original numerical model) in parallel using the 100 selected samples. The model's output will be the cumulative concentration across 7 distinct liquid-to-solids ratios.
    h) "Update Training Set" by adding a new set of data obtained from the ADE-IPD model to the current training set for ongoing surrogate model improvement.
4. The "(Re)training Ensemble Surrogate Model" integrates each optimized surrogate model based on each VRD using both stacking by a random forest and linear interpolation by inverse-distance weighting, provided they fulfill the stopping criteria. Finally, the procedure to "Check Ensemble Surrogate Model's Stop Criteria" was implemented to terminate the integration of optimized surrogate models and to assess the ensemble model's accuracy relative to the original model. For this assessment, a test set of 40 samples, derived from Latin Hypercube sampling, was employed.

Design of Experiments (DoE)

The DoE refers to the methodical planning and generation of a sample set that effectively captures the fundamental characteristics of the input parameter space. This sample set is subsequently employed as the initial training data for the surrogate model. The model outputs associated with these samples are obtained by evaluating the expensive numerical model (i.e., ADE-IPD).

Table 3.1 presents the model parameters with 11 variables in total along with their corresponding ranges, utilized in this study to evaluate the ADE-IPD model (see Table C.1

for the constant values used in the ADE-IPD simulations). In this study, we employed two specific sets of initial samples to assess how the size of the initial sample and the adaptive sampling strategy affect the performance of individual surrogate models. Each individual surrogate model was initiated with the same sets of initial samples: one set containing 110 samples and another containing 550 samples. These sizes correspond to 10 and 50 times the number of parameters, respectively, and were both generated using Halton Sequence sampling.

Halton Sequences belong to a group of low-discrepancy sequences utilized for creating designs that effectively fill parameter space (Halton, 1960). A notable advantage is that the total number of data points does not need to be determined a priori, which is in contrast to conventional pseudo-random sequences and Latin Hypercube sampling (Allgeier & Cirpka, 2023).

It's worth noting, as demonstrated by Finkel, Grathwohl, 2017, that intraparticle porosity (ε) and particle diameter (d) exhibit the same sensitivity while reverse effect on the intraparticle-diffusion model. Therefore, the ratio $\frac{d}{\varepsilon}$ is utilized as a training parameter of the surrogate model, rather than the individual values.

Parameter [Unit]	Range	Description
Particle Diameter [mm]	$d_{\text{coarse}} \sim \mathbf{U}(0.2, 2)$ $d_{\text{intermediate}} \sim \mathbf{U}(0.02, 0.2)$ $d_{\text{fine}} \sim \mathbf{U}(0.002, 0.02)$	determines the size of particles in the sample
Volume Fraction [%]	$\frac{f_{\text{m}_{\text{coarse,intermediate,fine}}} \sim \mathbf{U}(0, 100)}{\sum(f_{\text{m}_{\text{coarse}}}, f_{\text{m}_{\text{intermediate}}}, f_{\text{m}_{\text{fine}}})=100}$	represents the proportion of each grain size in the mixture
Distribution Coefficient [L kg^{-1}]	$K_{\text{d}_{\text{coarse,intermediate,fine}}} \sim \mathbf{U}(0.1, 25)$	governs solute partitioning between solid and liquid phases
Intraparticle Porosity [-]	$\varepsilon_{\text{coarse,intermediate,fine}} \sim \mathbf{U}(0.01, 0.1)$	internal void space within particles
Initial Concentration per Mass Dry Solid [mg kg^{-1}]	$C_{\text{s}_{\text{ini}}} \sim \mathbf{U}(0.01, 5000)$	expresses initial mass of contaminant
Dispersivity [m]	$\alpha \sim \mathbf{U}(0.015, 0.06)$	determines longitudinal mixing within the mobile pore space

Table 3.1: Prior parameter distribution of the ADE-IPD model based on the German standard for column leaching tests (DIN 19528, 2009). **U**: Uniform distribution

Selection a New Virtual-Reality Dataset

In our investigation, each individual surrogate model was sequentially trained and optimized using a designated VRD. Two hundred VRDs were generated using Halton Sequence sampling, with each VRD representing a distinct parameter set (see Figure C.1 in the appendix for the VRDs coverage along the predefined parameter space). We explored how the selection order of VRDs, determined by different distance metrics, influences the convergence of final ensemble surrogate models. Two distinct selection techniques were employed. The first technique involved selecting a new VRD based on the minimum Mahalanobis distances between previously selected VRDs and the remaining datasets. Conversely, the second technique utilized the maximum Mahalanobis distances for VRD selection. It is important to note that the initial VRD was selected randomly.

The Mahalanobis distance is a metric for quantifying the distance between two data points in the space defined by relevant features (Mahalanobis, 2018). Its unique attribute lies in its capability to account for both unequal variances and correlations among features (De Maesschalck et al., 2000). As a result, it effectively measures distance by assigning

variable weights to the features of data points. In cases where features are uncorrelated and have identical uncertainty, the Mahalanobis distance converges to the Euclidean distance. The following describes the Mahalanobis distance $d(\boldsymbol{a}, \boldsymbol{b})$ between points $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{(\boldsymbol{a} - \boldsymbol{b})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{a} - \boldsymbol{b})}, \tag{3.12}$$

where Σ represents a scaling matrix. In the given context, $\boldsymbol{a}$ is the i-th VRD, ($\boldsymbol{\theta_i}$), and $\boldsymbol{b}$ is the mean of all previously used VRDs, ($\bar{\boldsymbol{\theta}}$), whereas Σ is the covariance matrix of all previously used VRDs.

Training of the Surrogate Models

After several tests (not shown), we found that employing tree-based models yields the utmost precision in emulating the original model's behavior. In this study, Extremely Randomized Trees (ExtraTrees; Geurts et al., 2006) were employed as the training algorithm during adaptive sampling for an individual surrogate model, chosen for their strong capability to capture the leaching behavior of contaminants (Ershadi et al., 2023). Within each adaptive-sampling loop, the random-search method (Bergstra & Bengio, 2012, see B.10 in the appendix for more detail) was employed for optimizing the surrogate model's hyperparameters (i.e., the number of estimators, the maximum depths, maximum features, and minimum samples split).

Model Validation and Stopping Criteria of Single Surrogate Model

Within the optimization of each surrogate model for a specific virtual-reality dataset, the adaptive sampling is potentially stopped according to two predefined criteria: (a) if the RRMSE reaches 0.1 (see Eq. 3.11) or (b) if the size of the training set achieves a maximum of 1100 samples (following the recommendations of Zhu et al., 2023, suggesting that model performance will unlikely further increase beyond sample numbers larger than 100 times the number of parameters).

Infill Criteria

Within the framework of optimization, adaptive sampling involves the iterative selection of new data points to enhance the accuracy and coverage of the surrogate model's parameter space in an optimized manner. The primary goal of adaptive sampling is to meticulously curate a training dataset to faithfully reflect the original model while minimizing the necessity for expensive model evaluations.

Within this context, each iteration of adaptive sampling, involves the selection of 100 samples from a substantial pool of 5,000 proposed samples, previously generated by the Halton Sequence. This selection adheres to predefined infilling criteria aimed at balancing exploration and exploitation. Specifically, 25 samples are chosen based on maximizing the EI (exploitation), 50 samples are selected to maximize the standard deviation (exploration), and the remaining 25 samples are identified based on their Mahalanobis distance being below a specified threshold relative to the current best parameter set (exploitation).

Exploration in this context entails probing new and uncertain areas within the parameter space, thereby preventing the surrogate model from prematurely converging on local optima. Conversely, exploitation focuses on regions where the surrogate model has pre-

viously identified promising results (i.e., regions with the smallest value of the objective function), with the aim of enhancing these findings further.

Maximizing Expected Improvement (EI): The expected improvement is a widely used criterion in adaptive-sampling strategies of surrogate models and Bayesian optimization. It efficiently guides the selection of new data points in the design space to evaluate how much improvement of the current surrogate model is expected if a new sample is obtained.

The general form of the EI function for a specific input parameter $\boldsymbol{\theta}$ is defined as:

$$EI(\boldsymbol{\theta}) = \begin{cases} (f_{\text{obj}}(\boldsymbol{\theta}_{\text{best}}) - f_{\text{obj}}(\boldsymbol{\theta}))\Phi(Z) + \sigma(\boldsymbol{\theta})\phi(Z) & \text{if } \sigma(\boldsymbol{\theta}) > 0, \\ 0 & \text{if } \sigma(\boldsymbol{\theta}) = 0. \end{cases} \tag{3.13}$$

$$Z = \frac{f_{\text{obj}}(\boldsymbol{\theta}_{\text{best}}) - f_{\text{obj}}(\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta})}, \tag{3.14}$$

where $f_{\text{obj}}(\boldsymbol{\theta})$ represents the surrogate model that approximates the objective function at point $\boldsymbol{\theta}$, $\boldsymbol{\theta}_{\text{best}}$ denotes the parameter set that minimizes the objective function, while $f_{\text{obj}}(\boldsymbol{\theta}_{\text{best}})$ is the current best-observed value of the objective function. Additionally, $\sigma(\boldsymbol{\theta})$ denotes the corresponding standard deviation, capturing prediction uncertainty. The functions $\Phi(Z)$ and $\phi(Z)$, represent the cumulative distribution function and the probability density function of the standard normal distribution, respectively.

Note that ExtraTrees differs fundamentally from probabilistic models such as Gaussian Processes. In the case of a single point, $\boldsymbol{\theta}$, the predictive distribution of an ExtraTrees model does not adhere to the normal distribution. Consequently, the estimation of the EI has been changed to:

$$EI(\boldsymbol{\theta}) = \max(f_{\text{obj}}(\boldsymbol{\theta}_{\text{best}}) - f_{\text{obj}}(\boldsymbol{\theta}), 0). \tag{3.15}$$

We compute $EI(\boldsymbol{\theta})$ and select the 25 samples with the highest EI values:

$$\boldsymbol{\theta}_{EI_{\max}} = \arg\max_{\boldsymbol{\theta}} \{EI(\boldsymbol{\theta})\}, \quad |\boldsymbol{\theta}_{EI_{\max}}| = 25, \tag{3.16}$$

where $\boldsymbol{\theta}_{EI_{\max}}$ signifies the set of points where the EI is maximized, and $\arg\max_{\theta}$ denotes the argument, i.e., the value of $\boldsymbol{\theta}$, at which the EI function achieves its maximum. The condition $|\boldsymbol{\theta}_{EI_{\max}}| = 25$ explicitly indicates our interest in the top 25 points where the EI is maximized, guiding the optimization process toward samples with the greatest potential for enhancement in our analysis.

Maximizing Standard Deviation: Maximizing standard deviation entails prioritizing regions characterized by elevated uncertainty, particularly in scenarios with limited training data for a specific region, resulting in augmented uncertainty when predicting unobserved data points. Training the surrogate model in these uncertain regions enhances its robustness and reduces the risk of being trapped in a local minimum. The evaluation of input-space sparsity and the identification of suitable additional

data points with substantial prediction uncertainty are guided by the normalized standard deviation σ_n, also denoted coefficient of variation, as a primary criterion.

This criterion serves as a metric for the uncertainty associated with each individual tree within an ExtraTrees algorithm, providing insights into the variability and confidence levels linked to a specific prediction. The normalized standard deviation $\sigma_n(\boldsymbol{\theta})$ is determined by:

$$\sigma_n(\boldsymbol{\theta}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \bar{y}}{\bar{y}}\right)^2}, \tag{3.17}$$

where N represents the number of trees in a forest, y_i denotes the prediction of each tree, and $\bar{y}$ is the mean of all predictions within the forest. To prioritize samples with the highest uncertainty, we select the top 50 samples with the maximum normalized standard deviation:

$$S_{\max} = \arg\max_{\boldsymbol{\theta}} \{\sigma_n\}, \quad |S_{\max}| = 50, \tag{3.18}$$

where $\arg\max_{\theta}$ denotes the argument (or values) of $\boldsymbol{\theta}$ that maximizes this standard deviation. The cardinality $|S_{\max}|$ explicitly specifies that the selected set $S_{\max}$ contains 50 elements, prioritizing the samples with the maximum uncertainty in the context of our analysis.

Selection by the Mahalanobis Distance to the currently best parameter set:

In the context of our adaptive-sampling approach, we leverage the Mahalanobis distance to measure the separation between the proposed data points and the current best sample with the least RRMSE. By integrating this metric into the infill criteria, we significantly enhance the adaptability of our strategy, particularly within the trusted region for exploitation.

For a proposed sample vector $\boldsymbol{\theta}$ in a multivariate space with the current best sample vector $\boldsymbol{\theta}_{\text{best}}$ and covariance matrix of the training set Σ, we use the Mahalanobis distance $d(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{best}})$ according to Eq. (3.12) as selection criterion.

During each iteration of the adaptive-sampling loop, multiple proposed samples are generated based on a multivariate normal distribution, where the mean and covariance matrix are derived from the latest training set in the adaptive-sampling loop. The generation of the proposed samples is facilitated by the Cholesky decomposition of the covariance matrix:

$$\boldsymbol{\theta}_i = \bar{\boldsymbol{\theta}} + \mathbf{L}\boldsymbol{\xi}_i \tag{3.19}$$

$$\text{with } \Sigma = \mathbf{L}\mathbf{L}^{\mathrm{T}}, \tag{3.20}$$

where $\boldsymbol{\theta}_i$ and $\bar{\boldsymbol{\theta}}$ are the random sample i and the mean of the parameter vector $\boldsymbol{\theta}$, $\mathbf{L}$ is the lower triangular matrix of the Cholesky decomposition, and $\boldsymbol{\xi}_i$ is a vector of random values with the same size as $\boldsymbol{\theta}$ drawn from a standard normal distribution. Note that any other square-root of matrix Σ instead of the Cholesky decomposition

would do (e.g., a matrix with identical eigenvalues but square-rooted eigenvalues as Σ).

To exclusively exploit the trusted region, we randomly choose 25 samples from a pool of 10,000 proposed samples generated from the multivariate normal distribution outlined above, focusing on those with Mahalanobis distances below a predefined threshold. This threshold is determined as follows:

$$threshold = (d_{\min}) \times \left(1 + \frac{f_{\text{obj}}(\boldsymbol{\theta}_{\text{best}})}{counter}\right), \tag{3.21}$$

where $(d_{\min})$ represents the minimum Mahalanobis distance among the proposed samples, $f_{\text{obj}}(\boldsymbol{\theta}_{\text{best}})$ is the observed objective value based on the current best sample, and *counter* is the current number of adaptive-sampling loop iterations. By this procedure, the range of investigation dynamically adjusts based on both the accuracy of the current surrogate model and the iteration number in the adaptive-sampling process.

3.2.3 *Ensemble Model*

Ensemble modeling combines the capabilities of several surrogate models to augment predictive precision. In this study, we compare two distinct types of ensemble models: stacking using a random forest for the model combination and linear interpolation by inverse-distance weighting. Both techniques combine the predictions of multiple base-models (i.e., multiple surrogate models), recognizing that different (surrogate) models exhibit different distinct strengths and weaknesses (Wolpert, 1992; Breiman, 1996). Through the amalgamation of their predictions, better overall performance can often be attained.

As *base-models* we use a set of trained individual surrogate models based on distinct virtual-reality datasets, specifically constructed using the ExtraTrees algorithm (Geurts et al., 2006). Each trained surrogate model predicts a set of cumulative concentrations relevant to the considered input set. The *meta-model* then combines the individual surrogate models:

$$\hat{y}_{\text{stacked}} = f_{\text{meta}}\left(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\right), \tag{3.22}$$

where $\hat{y}_{\text{stacked}}$ is the final prediction made by the meta-model, and $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are the predictions of the base-models. f_{meta} represents the function that combines the individual surrogate models, and n signifies the total number of individual surrogate models included in the ensembling process.

When we refer to *model stacking*, the meta-model consists of a random forest (Breiman, 2001), whereas *inverse-distance weighted* (IDW) interpolation refers to a linear combination of surrogate model outputs:

$$\hat{y}_{\text{IDW}} = \sum_{i}^{n} w_i \hat{y}_i, \tag{3.23}$$

where w_i is the weight of the surrogate model i computed by the inverse Mahalanobis distances of the surrogate models from the test sample as defined below:

$$w_i = \frac{\frac{1}{d(\boldsymbol{\theta_i}, \boldsymbol{\theta}_{\text{test}}) + \epsilon}}{\sum_{j=1}^{n} \frac{1}{d(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{\text{test}}) + \epsilon}}, \tag{3.24}$$

where ϵ is a small scalar to avoid division by zero. Instead of inverse-distance weighting, we could also apply Gaussian Process Regression or any other linear estimator that ensures a larger influence of a given surrogate model for closer proximity of the parameter set to the test sample used to construct the surrogate model. In this context, it is important to note that the Mahalanobis distance applied in Eq. (3.24) considers the distance of each surrogate model $\boldsymbol{\theta_i}$ to the test point $\boldsymbol{\theta}_{\text{test}}$ rather than to the best value $\boldsymbol{\theta}_{\text{best}}$ or the mean $\bar{\boldsymbol{\theta}}$, as used above.

To ensure the resilience of ensemble surrogate models across the entire parameter space, we evaluated ensemble surrogate models based on random forest stacking and inverse-distance weighted interpolation using a test set composed of 40 sample sets. The process of ensembling individual surrogate models will continue until the objective function value is reduced to 0.1 over the test set (refer to Eq. 3.11).

3.2.4 *Model Calibration*

Model calibration refers to the process of adjusting model parameters to minimize the discrepancy between model predictions and experimental data. Various methodologies and techniques exist for model calibration, each with its strengths and limitations. The advantage of Bayesian techniques is that they determine the full parameter distribution conditioned on the data rather than a single best estimate. In this study, we employ NPE, drawing from the observations of column leaching tests as reported by Naka et al., 2016 (refer to Figure 1.7 for the flowchart of simulation-based inference).

The primary objective in NPE is to infer the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x_o})$ given a specific observation $\boldsymbol{x_o}$. In NPE, the forward model (in our case the ensemble surrogate model) is initiated across various parameter values sampled from a prior distribution $p(\boldsymbol{\theta})$ (Moss et al., 2023). Then, a neural density estimator $q_\phi(\boldsymbol{\theta}|\boldsymbol{x})$ (where q_ϕ is known as a normalizing flow; Papamakarios et al., 2019; Kobyzev et al., 2021) with learnable parameters ϕ is trained by minimizing the expected negative log-probability on simulated data according to (Deistler et al., 2022):

$$\min_{\phi} \mathcal{L} = \min_{\phi} \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta})} \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})} [-\log q_\phi(\boldsymbol{\theta}|\boldsymbol{x})]. \tag{3.25}$$

The loss $\mathcal{L}$ is formulated to minimize the expected negative log-likelihood of the model. This expectation is computed over two distributions: first, the prior distribution $p(\boldsymbol{\theta})$ of the model parameters $\boldsymbol{\theta}$, and second, the conditional distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ of the simulated observations $\boldsymbol{x}$ given the parameters $\boldsymbol{\theta}$.

The term $-\log q_\phi(\boldsymbol{\theta}|\boldsymbol{x})$ within the expectation denotes the negative log-likelihood of observing $\boldsymbol{x}$ given $\boldsymbol{\theta}$, estimated by the neural density estimator $q_\phi(\boldsymbol{\theta}|\boldsymbol{x})$ parameterized by ϕ. The goal is to minimize this quantity across all possible combinations of $\boldsymbol{\theta}$ and $\boldsymbol{x}$, sampled from their respective distributions. This formulation aims to train the model to better approximate the true data distribution and improve its predictive performance.

Through this process, the density estimator directly approximates the posterior distribution (Papamakarios & Murray, 2016; Lueckmann et al., 2017). In the model calibration

phase, NPE is carried out utilizing the flexible interface of the SBI toolbox in Python developed by Tejero-Cantero et al., 2020.

3.3 RESULTS AND DISCUSSION

3.3.1 *Verification of the Surrogate Model*

Surrogate Models' Optimization

In this section, we first assessed the effectiveness of optimizing individual surrogate models within the ARS loop. We specifically focus on examining the impact of the DoE on the optimization process. This analysis is conducted for two scenarios: one utilizing the DoE with 110 initial samples, and the other with 550 initial samples. Figure 3.5A presents the distribution of RRMSE for all 200 surrogate models in each adaptive-sampling iteration, utilizing training samples characterized by low numerical error (e.g., <5%). Please note that for the DoE with 550 initial samples, the adaptive-sampling iteration is only carried out until iteration 6, because this is when the stopping criterion of 1100 training samples is reached. Despite this difference in the number of iterations, both scenarios show a progressive reduction in predictive error facilitated by the adaptive-sampling process. As expected, starting with 550 initial samples involves a better initial performance of the surrogate models. This, however, does not prevail over several iterations of adaptive sampling.

Figures 3.5B-C provide additional insight into the influence of the initial number of training samples on the performance of individual surrogate models: smaller initial sample sizes (i.e., 10 times the number of parameters as suggested by Jones et al., 1998) can help prevent the initial surrogate model from becoming overly tailored to specific dataset peculiarities and instead promote the discovery of generalizable trends. By providing smaller but still representative initial training data, surrogate models have a better chance to reach a higher accuracy for each specific VRD through the adaptive-sampling loop. This approach leads to faster convergence toward the predefined accuracy stopping criteria, as demonstrated in Figure 3.5B. This figure shows that surrogate models with 110 initial samples reach the specified accuracy threshold more rapidly and with fewer total training samples compared to those with 550 initial samples. However, it's important to acknowledge that an excessively low number of initial samples can lead to underfitting, where the model fails to capture essential patterns in the data, potentially resulting in poor performance. Thus, finding an optimal balance in the size of the initial dataset is crucial for achieving effective model performance.

Furthermore, the additional ARS iterations offer more opportunities for the models with 110 initial samples to learn from new data points and refine their predictive capabilities. The iterative nature of the adaptive-sampling process allows surrogate models with 110 initial samples to gradually adjust and converge towards more accurate representations of the underlying objective function, as evidenced by decreasing RRMSE values (Figure 3.5C).

Due to the lower RRMSE of the surrogate models with 110 over 550 initial samples, we have opted to utilize the surrogate models with 110 initial samples for the remainder of the study.

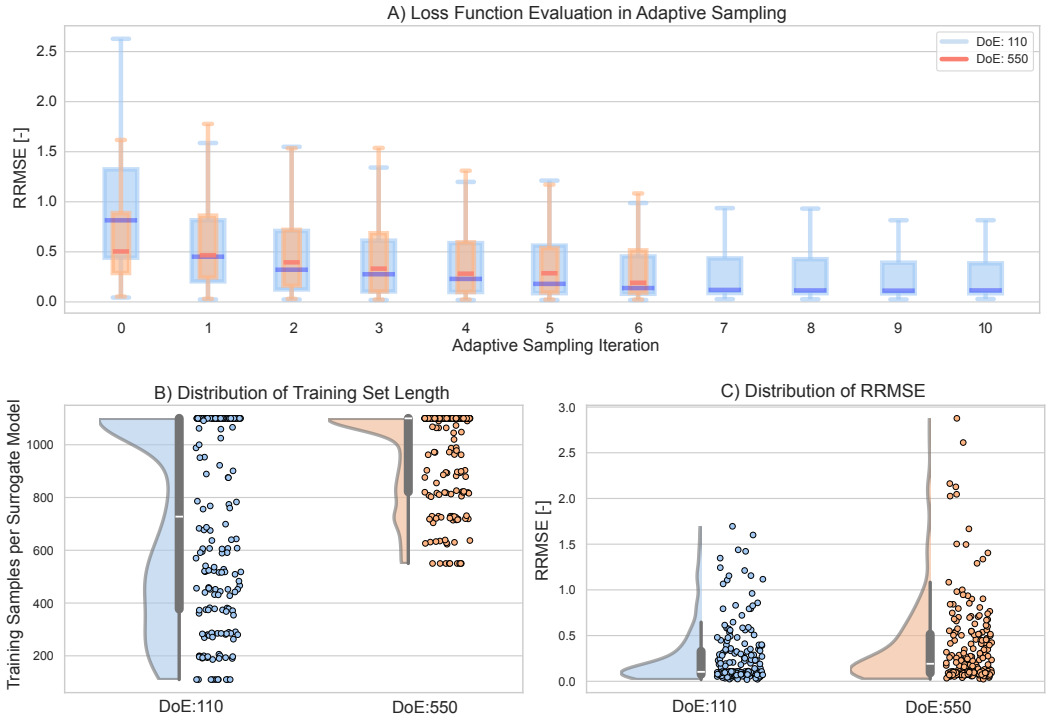


Figure 3.5: Optimizing surrogate models through adaptive sampling: (A) Boxplots illustrate the Relative Root Mean Squared Error (RRMSE) for 200 surrogate models across adaptive-sampling iterations, utilizing Design of Experiment (DoE) with initial sample sizes of 110 and 550. These boxplots provide a clear visual comparison of the model performance over the iterations, showcasing the reduction in RRMSE as the sampling progresses. (B) Half-violin plots, embedded with data points, represent kernel density estimates of the full distributions of the number of training samples used to train each individual surrogate model upon reaching the stopping criteria. Each data point corresponds to a specific surrogate model, providing insights into the variation in training set sizes between the two initial Design of Experiments (DoE) sample sizes of 110 and 550. (C) The half-violin plots illustrate the distribution of RRMSE values after the stopping criteria have been met for each of the 200 surrogate models based on initial DoE sample sizes of 110 and 550. The enclosed box within each half-violin illustrates key statistical measures, including the minimum, first quartile, median (highlighted by a white dash), third quartile, and maximum data values.

General Performance of Ensemble Surrogate Models

The performance of ensemble surrogate models depends on the selected ensembling technique. The two proposed techniques were evaluated by means of a test set of 40 random samples generated by the numerical model.

For random forest stacking of surrogate models, our findings demonstrate that selecting virtual-reality datasets based on the maximum Mahalanobis distances facilitated faster convergence of the ensemble surrogate model towards the predefined stop criteria, achieving this after stacking 118 individual surrogate models (refer to Table C.2). This enhanced convergence rate can be attributed to the model's ability to explore new areas and incorporate learning from these regions. In contrast, stacking strategies based on minimum Mahalanobis distances were found to enhance predictions only within specific regions, potentially limiting their effectiveness for broader predictive tasks across the entire predefined parameter range, thereby affecting their generalizability. It is important to note that

for the subsequent analyses, we utilized the stacked model that achieved quicker convergence, specifically the one employing the maximum Mahalanobis distances.

The comparative analysis demonstrates a clear advantage of the random forest stacking ensemble over the inverse-distance weighting (IDW) interpolation method (see Figure 3.6). The random forest stacking method consistently achieves a significantly lower RRMSE of 0.09 compared to 0.40 for the IDW method, underscoring its higher predictive accuracy. Moreover, the random forest stacking method shows a smaller bias, as expressed by a mean of the relative residuals (4.59% versus -29.78%), a narrower range of the residual distribution, as reflected by a lower standard deviation (9.22% versus 34.44% for IDW) and a near-zero skewness (-0.07 versus 0.76). All these metrics confirm a superior predictive precision and reduced systematic error of the random forest stacking in comparison to inverse-distance weighting.

The advantage of random forest stacking likely arises from its ability to combine the base models in a non-linear manner. This suggests a potentially non-linear relationship between the original model's inputs and outputs, which cannot be reproduced by linear interpolation. Moreover, through training a meta-model on the predictions of the surrogate models, the stacking approach facilitates error mitigation and harnesses the diversity inherent in the base models. In comparison, inverse-distance weighting interpolation appears to be more vulnerable to propagating errors from underperforming individual surrogate models, as it lacks a comparable corrective mechanism (Shams et al., 2021). Figure 3.6B illustrates this effect by demonstrating a bias towards higher predicted values, leading to a higher RRMSE. We may not exclude, however, that other linear interpolation techniques might perform better than inverse-distance weighting interpolation.

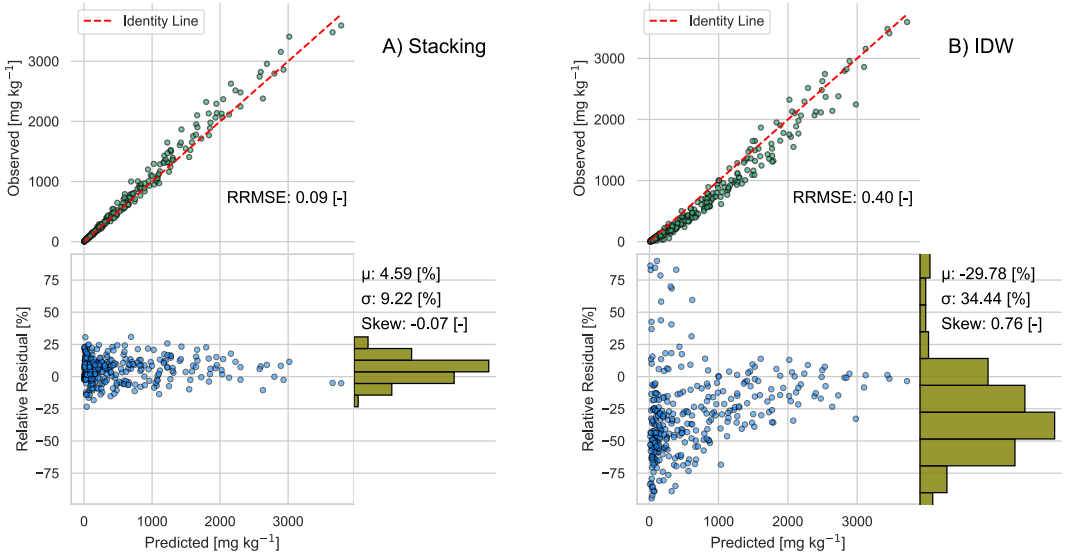


Figure 3.6: Performance of the ensemble surrogate models for predicting cumulative concentrations in soil samples at 7 different liquid-to-solids ratios (0.1, 0.2, 0.5, 1, 2, 5, 10 [L kg^{-1}]): (A) Random forest stacking and (B) inverse-distance weighting (IDW) interpolation, both evaluated on the same predefined test set. Blue scatters depict the relative residuals $\frac{\text{observation - prediction}}{\text{observation}}$ between actual and predicted cumulative concentrations. The histogram illustrates the distribution of relative residuals with statistical metrics for the mean (μ), standard deviation (σ), and skewness (Skew) of these residuals. The green scatters showcase actual vs. predicted values, with the red dashed line (identity line) indicating perfect predictions. The prediction accuracy of the surrogate models is quantified using the Relative Root Mean Squared Error (RRMSE).

3.3.2 *Analysis of Posterior Distributions*

For model calibration using simulation-based inference, the neural network in the training phase of NPE was trained on a dataset comprising $n_{\text{simulation}} = 10,000$ surrogate model simulations sampled from the prior distribution. Subsequently, 50,000 points were sampled from the trained neural posterior network based on the observed data. In the following, the original and surrogate models undergo evaluation on 100 random samples drawn from the posterior distribution to derive their respective predictive distributions.

Figure 3.7A shows the marginal distribution of each parameter. The prior distribution was sampled from uniform parameter distributions set according to the German standard for column leaching tests (DIN 19528, 2009; see Table 3.1). We used neural posterior distribution to obtain the posterior parameter distribution for the observed cumulative copper concentration in the column experiments of Naka et al., 2016 with two different soils A and B. Figure 3.7A shows several notable features of the marginal posterior distributions. For example, the initial concentration of contaminant on the dry solid ($C_{s_{\text{ini}}}$) exhibits the narrowest parameter distribution in both soils, demonstrating its strong influence on model predictions. The high sensitivity of the surrogate model to variations of $C_{s_{\text{ini}}}$ is expected, as this parameter directly controls the final (i.e., long-term) cumulative concentration. Also, the K_{d} values in all fractions exhibit high sensitivity; however, their posterior distributions are not as narrow as those of $C_{s_{\text{ini}}}$, leading to higher posterior uncertainty in these parameters. This is due to the negative correlations between each of these K_{d} values in different fractions, as illustrated in Figure C.2 in the supporting information. In soil B, these corre-

lations are more pronounced. Moreover, since the majority of the mass fraction in soil B is associated with the coarse fraction, $K_{\mathrm{d}_{\mathrm{coarse}}}$ exhibits greater sensitivity and leads to narrower posterior distributions than $K_{\mathrm{d}_{\mathrm{inter}}}$ and $K_{\mathrm{d}_{\mathrm{fine}}}$. However, in soil A, we observe a nearly identical shape of the posterior distribution for K_{d} values due to the fact that the mass fraction is almost uniform across these three fractions. The dispersivity and the ratio of diameter over intraparticle porosity for all three fractions exhibit the smallest sensitivities in predicting cumulative concentration across both soils.

To evaluate the effectiveness of reproducing laboratory measurements, 100 random samples from the posterior parameter distribution were applied to both the surrogate and the numerical model. Figures 3.7B-C depict predictions generated from these samples, showcasing a perfect fit for both the surrogate and the original models. Additionally, the confidence intervals exhibit almost similar sensitivity of both models.

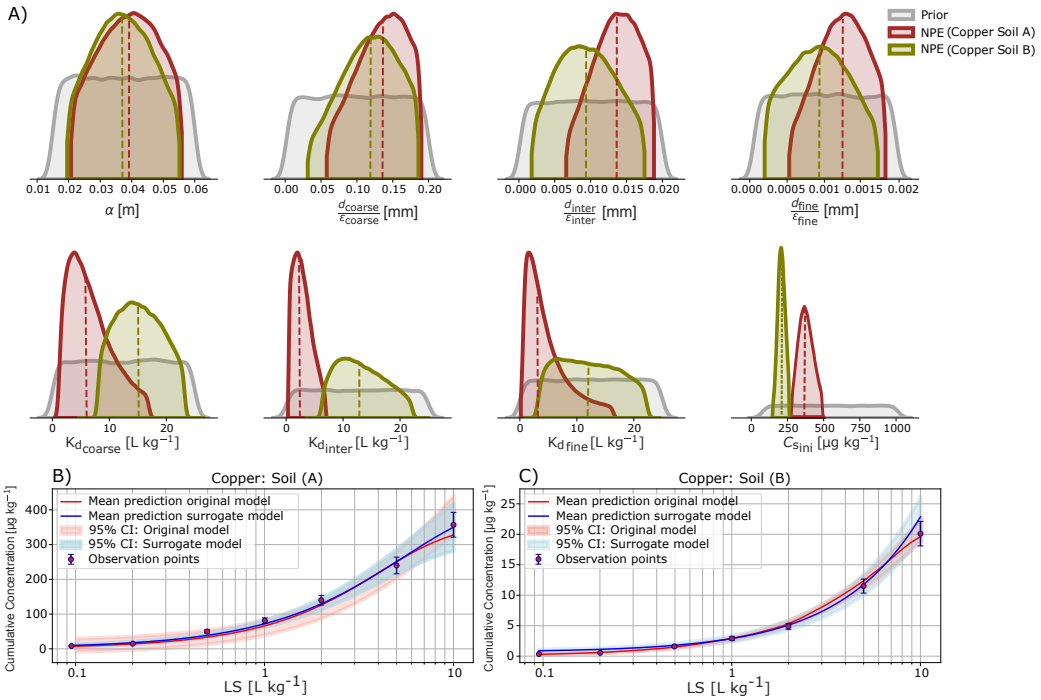


Figure 3.7: Comparison of model calibration for copper in soil samples A and B using observation data conducted by Naka et al., 2016. (A) Prior and posterior parameter distributions using Neural Posterior Estimation (NPE) for the original model and the stacking surrogate model. Densities are cut off at the 94% highest density intervals, and vertical dashed lines indicate the mean parameter values. The subscripts 'coarse', 'inter', and 'fine' denote the fractions with coarse, intermediate, and fine particle diameters, respectively. The color scheme used in the graphical representations is as follows: grey denotes the prior parameter distributions, while red and green indicate the posterior distributions of parameters for soil samples A and B, respectively. (B-C) Comparison of model predictions (mean and 95% confidence intervals) from the original model and the stacking surrogate model using 100 samples from the posterior parameter distribution.

3.4 CONCLUSIONS

In this study, we have developed two ensemble surrogate models for column leaching tests across the entire defined parameter range according to the German standard (DIN 19528, 2009). The original numerical model simulates one-dimensional advective-dispersive transport coupled to intraparticle diffusion using a cell-centered Finite Volume Method. The ensemble surrogate models accelerate the simulations by a factor of over 1000 in comparison to the original model. On average, the original model takes more than 150 seconds, whereas the stacking surrogate model requires only 0.1 seconds (refer to B.8 for details on the computer system used to develop the model). This acceleration facilitates the rigorous interpretation of column leaching tests in short computing times while maintaining a reasonable level of accuracy.

Each base surrogate model underwent optimization utilizing the adaptive-recursive sampling. Our investigation included an analysis of the impact of multi-infill criteria (maximizing expected improvement and standard deviation and keeping a minimizing Mahalanobis distances) with the trade-off between exploration and exploitation during adaptive sampling. Our analysis indicates that the initial sample size affects surrogate models' potential to achieve high accuracy. Specifically, a median RRMSE of 0.11 could be achieved with an initial sample size of 110, i.e., 10 times the number of parameters (compared to RRMSE = 0.19 for an initial sample size of 550) (see Figure 3.5). However, determining the optimal number of initial samples remains a vital consideration. Moreover, further analysis of the effect of the proposed and training sample sizes on model performance could provide valuable insights for optimizing surrogate model development.

We effectively applied the random forest stacking ensemble surrogate model instead of the costly original model in the inference of parameters from measurements of the cumulative concentration of copper from leaching tests conducted on two distinct soil types by employing neural posterior estimation (NPE) on 10,000 simulations of the ensemble surrogate model. Drawing samples from the estimated posterior distribution and running the full model, we could accurately reproduce the measured data (Figure 3.7), indicating that the posterior distribution of parameters was indeed conditioned on the measurements.

It is important to acknowledge that the ensemble surrogate models developed here are tailored specifically to simulate the column leaching test of contaminants from soil, adhering to German standards (DIN 19528, 2009). Consequently, they may not account for variations in other standards, such as FprCEN/TS 16637-2 (DIN CEN/TS 16637-2, 2014), and US EPA test method 1314 (US EPA Method 1314, 2013) which incorporate features like stop flow in their process-based models for simulating column leaching tests. Additionally, it's worth noting that the parameter investigation ranges for certain parameters, like K_d, primarily focused on scenarios with fast to moderate leaching rates. Consequently, we have not considered scenarios involving slow leaching rates with high K_d, such as those associated with high molecular weight polycyclic aromatic hydrocarbons (PAHs). Note that the considered K_d range covers a large spectrum of soil types and environmental conditions depending on the compound considered. For selected heavy metals, this includes mainly mineral and acidic soils (see C.1 and Figure C.3), while for selected organic compounds, it can encompass both mineral and organic soils based on the organic carbon partition coefficient (K_{oc}) of the compound (see C.2 and Figure C.4).

4

MODELING PFAS SORPTION IN SOILS USING MACHINE LEARNING

Amirhossein Ershadi[1], **Joel Fabregat-Palau**[1], **Michael Finkel**[1], **Anna Rigol**[2,3], **Miquel Vidal**[2], **Peter Grathwohl**[1]

[1]*Department of Geosciences, University of Tübingen, Schnarrenbergstraße 94-96, 72076 Tübingen, Germany*
[2]*Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Martí i Franquès 1-11, Barcelona 08028, Spain*
[3]*Institut de Recerca de l'Aigua (IdRA), Universitat de Barcelona, Martí i Franquès 1-11, Barcelona 08028, Spain.*

Published in the **Environmental Science and Technology**
https://doi.org/10.1021/acs.est.4c13284

Author Contributions

First author:
Scientific ideas: 50%, Data processing: 50%,
Analysis & interpretations: 50%, Paper writing: 40%

4.1 INTRODUCTION

Per- and polyfluoroalkyl substances (PFAS) are anthropogenic compounds characterized by high environmental persistence due to their strong C–F bonds (Prevedouros et al., 2006). Currently, PFAS are defined as substances that contain at least one fully fluorinated methyl or methylene carbon atom (without any H/Cl/Br/I atom attached to it) (Wang et al., 2021). These compounds are broadly classified as perfluorinated, where all hydrogen atoms are replaced by fluorine, or polyfluorinated, where only some hydrogens are replaced. PFAS can also be categorized into short- and long-chained compounds based on the number of fluorinated carbons, with short-chain perfluorosulfonic acids (PFSA) and perfluorocarboxyl acids (PFCA) having less than six and seven fluorinated carbons, respectively (Buck et al., 2011). PFAS represent a diverse chemical class with extensive industrial and consumer applications, including aqueous fire-fighting foams (AFFF), electronics, construction materials, and coatings for paper products (Glüge et al., 2020; Langberg et al., 2021). Anionic-type PFAS such as PFCA and PFSA have been widely studied for their occurrence, sorption, and toxicity (Li et al., 2018; Brusseau et al., 2020; Fenton et al., 2021), but neutral, cationic, and/or zwitterionic PFAS may constitute the majority of PFAS at certain AFFF-contaminated sites (Liu et al., 2022b; Schüßler et al., 2024). The sorption behavior of some of these PFAS has been overlooked, albeit current regulatory restrictions consider PFAS as a whole class (Bălan et al., 2021).

PFAS contamination is widespread, with concentrations in soils ranging from a few pg g^{-1} in remote areas to hundreds of μg g^{-1} at impacted sites (Rankin et al., 2016; Brusseau et al., 2020). Soils act as both filters and sources for PFAS contamination in groundwater (Röhler et al., 2021). Sorption parameters, such as the solid-liquid distribution coefficient (K_d), are essential for modeling contaminant transport in aquifers under saturated conditions. In unsaturated soils, however, the air-water distribution coefficient (K_{aw}) should also be considered, as it may lead to the retention of contaminants at the air-water interfaces (Guo et al., 2020; Brusseau, 2023). Therefore, understanding PFAS sorption in environmental matrices is essential for mitigating associated environmental concerns by informing risk assessment, remediation strategies, and regulatory decision-making. K_d values may be estimated from organic-carbon normalized sorption coefficients (K_{OC}) or by prediction models that account for certain soil and PFAS properties (Higgins & Luthy, 2007; Card et al., 2017; Knight et al., 2019; Fabregat-Palau et al., 2021; Umeh et al., 2021; Xie et al., 2024). These tools have limitations, such as their narrow PFAS range and reliance on oversimplified assumptions that overlook complex PFAS-soil interactions. Moreover, studies assessing the predictive performance of existing models, as well as a systematic analysis of PFAS features affecting sorption (e.g., the effect of PFAS chain length and functional groups), are lacking.

The compilation of K_d datasets enriched with both PFAS and soil characteristics by examining literature studies is critical for modeling purposes (Ma et al., 2023; Xie et al., 2024; Fabregat-Palau et al., 2024). Machine learning (ML) has emerged as a powerful method for modeling complex, non-linear relationships within large, multidimensional datasets to uncover patterns and correlations that conventional models might overlook (Ma et al., 2023; Moghadasi et al., 2023; Ershadi et al., 2023, 2024). Recently, Xie and coworkers developed an ML model on a literature-based K_d (PFAS) dataset consisting of 2,328 entries for 26 different anionic PFAS (Xie et al., 2024). Acceptance data criteria considered K_d values for the same PFAS/soil pairs at different pH conditions, especially those originating from a single

study (Nguyen et al., 2020). However, the speciation of ionizable PFAS with pK_a values within environmentally relevant pH ranges (e.g., perfluorosulfonamides (FOSA), $pK_a \approx 6$) was not considered (Rayne & Forest, 2009).

In this study, we systematically compiled a comprehensive literature-based K_d PFAS dataset, incorporating 1,274 entries for 51 different PFAS, including anionic, neutral, cationic, and zwitterionic species. Using ML algorithms, we developed a novel model that integrates both PFAS- (i.e., molecular weight (MW), hydrophobicity, pK_a) and soil-specific (i.e., pH, texture, cation exchange capacity (CEC), organic carbon content (C_{org})) descriptors. The model demonstrated superior predictive performance compared to other existing tools and, when combined with location-specific soil data (e.g., EU LUCAS topsoil repository), enabled the generation of spatial K_d maps. The dataset and model are available through the online platform PFASorptionML (https://hydrogeochem.geo.uni-tuebingen.de/pfas).

4.2 MATERIALS AND METHODS

4.2.1 *PFAS considered in this work*

Sorption data for a total of 51 PFAS were included in the dataset. These data comprise C_2 – C_{14} PFCA (i.e., TFA, PFBA, PFPeA, PFHxA, PFHpA, PFOA, PFNA, PFDA, PFUnA, PFDoA, PFTeA and PFTrA) and C_4 – C_{10} PFSA, (i.e., PFBS, PFPeS, PFHxS, PFHpS, PFOS, PFNS, and PFDS), including a cyclic (i.e., PFEtCHxS) species. PFCA and PFSA have been widely detected in soils and water at varying concentrations (Xiao, 2017; Brusseau et al., 2020), and they are known for their resistance to biodegradation. Particular attention was given to TFA due to its high volatility and low sorption affinity to soils, resulting in elevated levels in the atmosphere ($\leq$ 7 ng m^{-3}), water ($\leq$ 3 µg L^{-1}), and soils ($\leq$ 2 ng g^{-1}) (Richey et al., 1997; Xie et al., 2020). The dataset also includes C_6 – C_{10} perfluorophosphonic acids (PFPA, i.e., PFHxPA, PFOPA, PFDPA) and C_{12} – C_{16} perfluorophosphinic acids ($C_{x/y}$ PFPiA, i.e., $C_{6/6}$ PFPiA, $C_{6/8}$ PFPiA and $C_{8/8}$ PFPiA). Both PFPA and PFPiA have been detected in water bodies and human serum samples, with concentrations ranging from 0.1 to 3.7 ng L^{-1} and 4 to 38 ng L^{-1}, respectively (Lee & Mabury, 2011; Xiao, 2017). Additionally, the dataset includes C_4–C_8 FOSA (i.e., FBSA, FHxSA, PFOSA, and Et-FOSA), as well as N-methyl and N-ethyl perfluorooctane sulfonamidoacetic acids (FOSAA, i.e., N-MeFOSAA and N-EtFOSAA) species. These PFAS have been detected in water bodies at concentrations ranging 2.5 – 5.8 ng L^{-1} (Huang et al., 2019). Emerging PFAS classes were also included, such as perfluoroether carboxylic acids (PFECA, i.e., GenX and ADONA) and chlorinated polyfluoroalkyl ether sulfonate (PFAES, i.e., 8:2 Cl-PFAES). GenX and 6:2 Cl-PFAES have been found in water bodies at high concentrations ($<$ 5 and $\leq$ 112 µg L^{-1}) close to fluorochemical facilities and wastewater treatment plant effluents, respectively (Xiao, 2017; Munoz et al., 2019). Furthermore, data for C_4 – C_{10} *n*:2 fluorotelomer alcohols (FTOH, i.e., 4:2, 6:2, 8:2 and 10:2 FTOH) and C_4 – C_8 *n*:2 fluorotelomer sulfonates (FTS, i.e., 4:2, 6:2 and 8:2 FTS) were included. FTOH are known degradation products of side-chain fluorinated polymers and have been found in air samples at $\leq$ 0.3 µg m^{-3} (Schlummer et al., 2013; Schellenberger et al., 2019), whereas FTS are compounds present in AFFF-formulations and have been found in impacted sites at $\approx$ 4 ng g^{-1} (Brusseau et al., 2020). Additionally, we included data for C_6 – C_{10} cationic and zwitterionic PFAS (i.e., 6:2, 8:2 and 10:2 FtSaB, 6:2 FtSaAm, PFOSB, PFOAAmS, PFOAB, AmPr-FHxSA and TAmPr-FHxSA), which are predominant species in AFFF-formulations and have been found in both AFFF-impacted soil

and groundwaters (Liu et al., 2022b; Schüßler et al., 2024). The relative distribution of PFAS subfamilies and compounds in the dataset is illustrated in Figure 4.1.

Physicochemical properties of the PFAS were obtained from available data or estimated using EPISuite and the PubChem repository. Specifically, EPISuite (KOWWIN method versions 1.68 and 2.00) was used to estimate water solubility (S) and organic carbon-normalized sorption coefficients (K_{OC}), respectively. Octanol-water partition coefficients (K_{OW}) were obtained from PubChem's XLogP3 (version 3.0, released 2021.10.14). Sorption of charged compounds may be highly dependent on the amount and type of charged species (Fabregat-Palau et al., 2024). To include PFAS speciation in the model, acidity dissociation constants (i.e., pK_a) were compiled. Although commonly studied PFCA and PFSA compounds have low pK_a values ($pK_a \approx 1$), resulting in their presence as anionic species under most environmental conditions, some other PFAS compounds such as FOSA and some betaines exhibit pK_a values more aligned with environmental conditions (i.e., $pK_a \approx 6$) (Rayne & Forest, 2009; Mejia-Avendaño et al., 2020). The pH-speciation diagrams for representative PFAS are detailed in D.1.

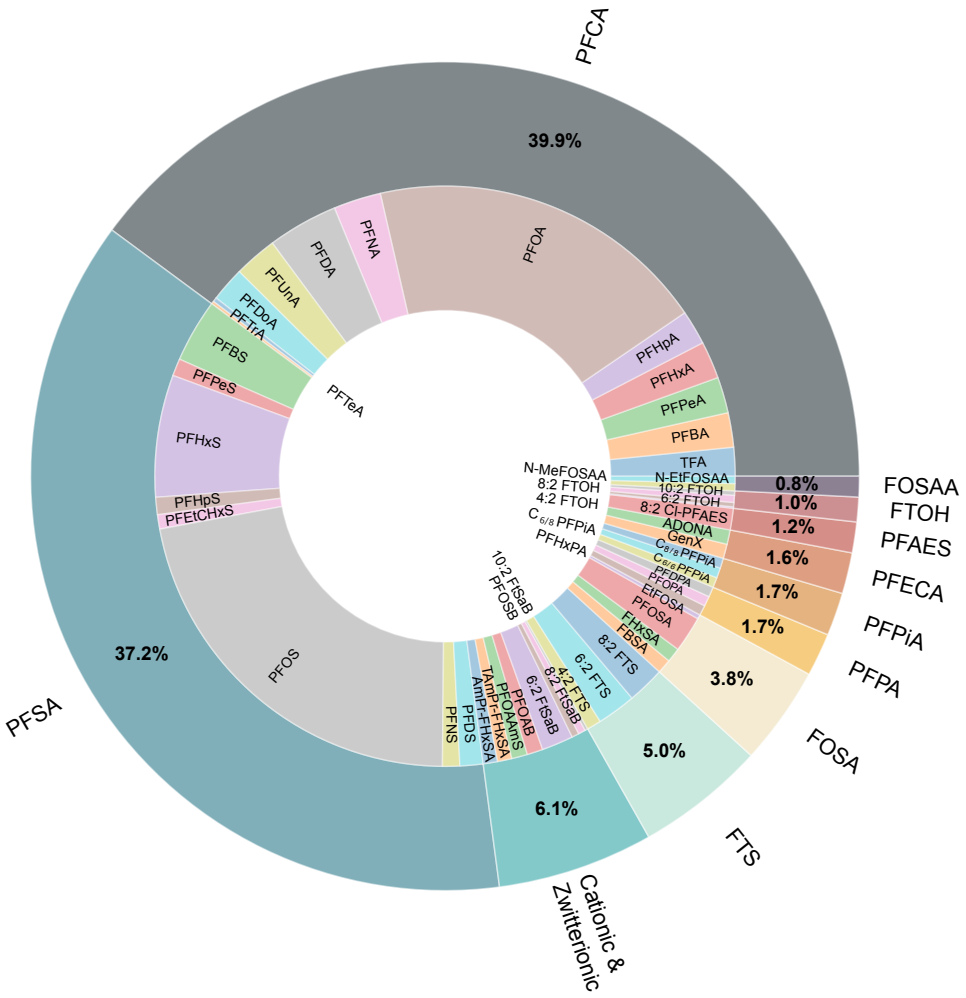


Figure 4.1: Relative distribution of PFAS subfamilies and respective compounds in the K_d (PFAS) dataset, including Perfluorocarboxylic Acids (PFCA), Perfluorosulfonamides (FOSA), Perfluorosulfonic Acids (PFSA), Fluorotelomer Sulfonates (FTS), Perfluorophosphonic Acids (PFPA), Fluorotelomer Alcohols (FTOH), Perfluoroalkyl Ether Carboxylic Acids (PFECA), Perfluorooctane Sulfonamidoacetic Acids (FOSAA), and Chlorinated Polyfluoroalkyl Ether Sulfonate (PFAES).

4.2.2 *Criteria for data compilation*

To be considered in the dataset, sorption data for PFAS must originate from a batch test following official guidelines meant to study sorption under saturated conditions at room temperature (i.e., $20 - 25^{\circ}C$), with slight variations (e.g., selection of contact solution) permitted (OECD, 2000). Data originated from other sources (e.g., K_d distribution originating from ratios of site-measured concentrations in both solids and liquid matrices) were not considered (Li et al., 2018). The dataset covers information on the data source and PFAS properties (i.e., MW, empirical formula, S, K_{OW}, and pK_a). We accepted sorption data reported on both soil and sediments, assuming analogous sorption behavior among these matrices in batch tests, but we excluded sorption data in pure organic substances (e.g., humic acids) or minerals (e.g., goethite, kaolinite) phases. The main physicochemical proper-

ties of the solids (i.e., pH, Fe and Al contents, CEC, C_{org}, and soil texture information (i.e., sand, silt, and clay contents)), as well as experimental details of the batch experiment (i.e., nature of the contact solution, solid-to-liquid ratio, initial PFAS concentration range) were additionally compiled for better understanding of sorption behavior.

K_d (PFAS) values were compiled as a representative sorption parameter. Straightforward reported K_d values, either originating from a low single-concentration spike sorption experiment (Nguyen et al., 2020; Fabregat-Palau et al., 2021) or derived from the lower concentration range of the sorption isotherm (Higgins & Luthy, 2006; Guelfo & Higgins, 2013), were considered. If several K_d (PFAS) values were reported from non-linear isotherm data at varying concentrations (Barzen-Hanson et al., 2017), the K_d value reported at the lowest concentration was preferred, as a better representative of PFAS environmental concentrations (Xiao, 2017; Brusseau et al., 2020). K_d (PFAS) values were similarly derived when needed from reported K_{OC} data (Campos Pereira et al., 2018; Campos-Pereira et al., 2022). Contrarily to other studies that accepted K_d (PFAS) values for the same PFAS/sorbent pair at different pH and/or ionic strength conditions (Xie et al., 2024), we adopted here stricter selection criteria. Only K_d values measured under conditions closely matching the original soil pH or at the lowest ionic strength were included (Nguyen et al., 2020; Cai et al., 2022). While this approach reduced the number of K_d entries in the dataset, it ensured a better representation of the original soil properties in the model.

In some cases, isotherm fitting parameters were reported, but no K_d (PFAS) values. Although some authors have estimated K_d values from isotherm parameters at a certain arbitrary concentration within the low isotherm range (Fabregat-Palau et al., 2023, 2024), in this study, we derived K_d (PFAS) data from isotherm fitted parameters (mainly resulting from Freundlich and, to a lesser extent, Langmuir fits) at the same PFAS concentration in water relative to the reported PFAS solubility (i.e., 10%), in agreement with other previous work (Kleineidam et al., 2002). Further details on the derivation of K_d (PFAS) values from literature studies are provided in D.2. Overall, 94 and 184 K_d (PFAS) entries originated from K_{OC} and isotherm fitted data (7% and 14% of the total entries, respectively). During the dataset construction, expert judgment was applied to identify individual entries that exhibited potential outlier behavior, defined here as unexpectedly high or low K_d (PFAS) values, particularly identified by examining individual K_d vs f_{OC} relationships (see Supplementary File 1 at https://hydrogeochem.geo.uni-tuebingen.de/pfas). Subsequently, we assessed entire PFAS compound classes for potential outlier trends, particularly by examining log K_{OC} vs log K_{OW} relationships. Table D.1 provides a summary of the number of entries and publications contributing data for each PFAS. Based on our assessment, 47 out of the 1,274 K_d (PFAS) entries (4% of the total) were classified as outliers, leaving a final dataset of 1,227 entries representative of 47 PFAS spanning 451 soil and sediment types across 47 studies.

4.2.3 *Model Development*

Feature Engineering

In developing the ML model, we carefully selected and engineered both PFAS- and soil-specific features. The soil-specific properties used in the model include soil pH [-], CEC [$cmol_+$ kg^{-1}], C_{org} [%], and textural (i.e., sand [%], silt [%], and clay [%] contents) information. These features were chosen based on their availability and established influence on

PFAS sorption. Other soil properties, such as iron (Fe) and aluminum (Al) contents, were excluded due to limited data availability (i.e., only 45% of the entries included available data), along with contradictory findings in the literature regarding their relevance to PFAS sorption (Oliver et al., 2019; Nguyen et al., 2020). Regarding soil pH, sorption of PFAS has been shown to decrease when increasing the pH of the batch contact solution for a given soil type (Nguyen et al., 2020). CEC is mainly the result of the soil organic matter content and, to some extent, of clay minerals (Tan & Dowling, 1984). While sorption of neutral and anionic PFAS in soils is known to be highly affected by C_{org} (Fabregat-Palau et al., 2021), sorption of cationic and zwitterionic organic compounds has been better related to CEC (Barzen-Hanson et al., 2017; Fabregat-Palau et al., 2024). Soil textural information was also included in the model, as clay minerals may impact PFAS sorption in scenarios of low C_{org} (Xiao et al., 2011; Fabregat-Palau et al., 2021).

Since the K_d (PFAS) dataset we compiled from literature contained gaps in soil physicochemical properties (i.e., pH, CEC, and soil texture), we constructed an additional dataset of selected soil properties (i.e., pH, CEC, C_{org}, and soil texture) using the SoilGrids repository (Poggio et al., 2021). This soil dataset consists of 2,039 entries distributed worldwide and was used to develop a K-nearest neighbor (KNN) imputer model (Troyanskaya et al., 2001), which allowed prediction of the soil property data gaps in the K_d (PFAS) dataset. This approach to addressing data gaps is an upgrade compared to other studies, where K_d modeling was restricted to the availability of specific soil properties (Fabregat-Palau et al., 2023, 2024), and has the potential to be implemented in further studies. Information on the KNN imputer model construction and implementation is provided in D.4.

PFAS-specific properties included in the model are MW [g mol^{-1}], log K_{OW} [-], and, for the first time, the molar net charge over MW ratio (i.e., the charge density [C g^{-1}]). MW was selected as a general descriptor of PFAS properties. Whereas K_{OW} was chosen as a good representative of the hydrophobic interaction between the PFAS alkyl chain and soil organic matter (Fabregat-Palau et al., 2021), the charge density is indicative of the electrostatic interaction between the PFAS functionalities and the charged soil surfaces. The net sign of this novel feature allowed inclusion of PFAS speciation in the model and differentiation of PFAS among cationic (charge density > 0), neutral or zwitterionic (charge density $\approx$ 0), and anionic (charge density < 0) species at the particular pH condition, as well as differentiation of the charge density among PFAS of the same subfamily (e.g., anionic PFBA and PFOA species). The abundance of each PFAS species was calculated according to the Henderson-Hasselbalch equation, ultimately providing information on the net charge of PFAS molecules in the system (Fabregat-Palau et al., 2024). For PFAS compounds containing a single ionizable group, the relative abundance of each species ($A_r(\text{PFAS})_i$) at a given soil pH was calculated using:

$$A_r(\text{PFAS})_{i=1} = \frac{1}{1 + 10^{(pK_{a1} - \text{pH})}} \tag{4.1}$$

$$A_r(\text{PFAS})_{i=2} = \frac{1}{1 + 10^{(\text{pH} - pK_{a1})}}, \tag{4.2}$$

where i denotes the different PFAS species at a certain pH condition (see D.1). For those PFAS that contained two ionizable groups, $A_r(\text{PFAS})_i$ was calculated as:

$$A_r(\text{PFAS})_{i=1} = \frac{1}{1 + 10^{(\text{pH} - pK_{a1})} + 10^{(2\text{pH} - pK_{a1} - pK_{a2})}} \tag{4.3}$$

$$A_r(\text{PFAS})_{i=2} = \frac{1}{1 + 10^{(pK_{a1} - \text{pH})} + 10^{(\text{pH} - pK_{a2})}} \tag{4.4}$$

$$A_r(\text{PFAS})_{i=3} = \frac{1}{1 + 10^{(pK_{a2} - \text{pH})} + 10^{(pK_{a1} + pK_{a2} - 2\text{pH})}}. \tag{4.5}$$

Whether these fractions were cationic, zwitterionic, anionic, or neutral was dependent on each PFAS pH-speciation diagram (see D.1). The molar net charge of each PFAS species ($z(\text{PFAS})_i$ [−]) was also dependent on the pH-speciation diagram (see D.1). Due to PFAS speciation under specific pH conditions, MW was recalculated to reflect the relative abundances of different species (e.g., accounting for protonated and deprotonated forms), thus defining an effective molecular weight at the specific pH condition. The charge density [C g^{-1}] was then determined as:

$$\text{Charge density} = \frac{\sum_{i=1}^{n} \left(A_r(\text{PFAS})_i \, z(\text{PFAS})_i\right) F}{\sum_{i=1}^{n} \left(A_r(\text{PFAS})_i \, \text{MW}(\text{PFAS})_i\right)}, \tag{4.6}$$

where F is the Faraday constant [C mol^{-1}].

The ranges of PFAS and soil properties used in the training set ultimately set the applicability boundaries of the model. A histogram of the distribution of entry values for each feature included in the model is shown in Figure D.2. The imputed soil property values follow the same distribution pattern as the original (i.e., raw) K_d (PFAS) dataset. MW of the selected PFAS ranged 114–770 g mol^{-1}, representative of PFAS with a fluorinated load (%F) ranging from 43 to 72% w/w. The charge density ranged from 188 to -854 C g^{-1}. Soil pH ranged from 2.8 to 9.0, and C_{org} ranged from 0.03 to 54%, and was skewed to those soils with low C_{org} (i.e., $< 2\%$). Individual examinations of the relative entry distribution across different C_{org} ranges were assessed for some PFAS with $N > 10$ entries (see Figure D.3) and indicated a lack of sorption data in organic soils. Soil textural fractions ranged from 0 to 100% for sand and silt, and 0 to 69% for clay, indicating that the soils cover a wide range of textural characteristics (see Figure D.4). CEC ranged from 0.1 to 140 $cmol_+$ kg^{-1}, and log K_d (PFAS) ranged from -1.4 to 3.95, thus spanning over five orders of magnitude.

K_d (PFAS) model construction and validation

To ensure that all features contributed equally to the learning process, each feature was normalized using min-max scaling, thus enabling a balanced contribution in the model (Patro & Sahu, 2015) (see details in A.4). Later, the data was randomly split using the holdout method, with 80% allocated for model training and 20% allocated for testing. This split ensures that the testing set remains independent from the training set, allowing for a better assessment of the model's performance (Hastie et al., 2009).

The predictive model for K_d (PFAS) was then constructed using a stacking ensemble approach (Wolpert, 1992), a two-layer framework that combines the strengths of multiple ML algorithms. The resulting model was termed as PFAS Sorption Stacking Model (PSSM). In the first layer, four base models—Ridge (Hoerl & Kennard, 1970), Random Forest (Breiman, 2001), Extremely Randomized Trees (Geurts et al., 2006), and Gradient Boosting (Friedman, 2001)—were trained on the input features $\mathbf{X}$, denoted as $\left(h_m(\mathbf{X}) \,\middle|\, m = 1, 2, \ldots, M = 4\right)$. These models generated a set of predictions $\hat{y}_m = h_m(\mathbf{X})$, which served as intermediate outputs for the subsequent layer. These predictions from the base models were then combined into a new dataset $\mathbf{Z}$, where $\mathbf{Z} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M\}$, to be used as input for the second layer (i.e., meta-model). In this study, the meta-model was a Multi-Layer Perceptron (Good-

fellow et al., 2016; Dubey et al., 2022). The meta-model learned to optimally combine the predictions from the base models, leveraging their collective strengths and generating the final output ($\hat{y}_{\text{final}}$) as shown in Equation 4.7:

$$\hat{y}_{\text{final}} = g(\mathbf{Z}) = g(h_1(\mathbf{X}), h_2(\mathbf{X}), \ldots, h_M(\mathbf{X})). \tag{4.7}$$

The predictive performance of the PSSM was evaluated using the normalized root mean squared error (NRMSE), the k-fold cross-validated normalized root mean square error (CV-NRMSE), and the ratio of performance to deviation (RPD) (see A.9 for additional information on these metrics).

Sensitivity Analysis for K_d (PFAS) Model

Understanding the sensitivity of a model to different input features provides insights into the robustness and reliability of the model and helps to explain model outputs. Two primary methods for sensitivity analysis in ML are SHapley Additive exPlanations (SHAP) (Shapley, 1953) and Partial Dependence (PD) (Greenwell et al., 2018). SHAP is a popular method based on cooperative game theory that provides model-agnostic, locally accurate explanations for individual predictions (Shapley, 1953). SHAP values quantify each feature's contribution by estimating the average marginal impact across all possible feature combinations (Sundararajan & Najmi, 2019). This approach ensures consistent, fair evaluations of each feature's influence across different combinations (Lundberg & Lee, 2017). The SHAP value (ϕ_i) for feature i is computed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right], \tag{4.8}$$

where F represents the full set of features, S is a subset of features excluding feature i, $|S|$ denotes the number of features in subset S, $f(S)$ represents the model prediction when only features in subset S are used, $f(S \cup \{i\})$ is the model prediction when feature i is added to subset S, and the term $\frac{|S|!(|F|-|S|-1)!}{|F|!}$ is a weighting factor that considers all possible permutations of feature subsets, ensuring that each feature's contribution is fairly evaluated across different combinations.

PD, in contrast, provides a global view of the relationship between a specific feature and the predicted outcome by averaging out the influence of all other features (Friedman, 2001). PD analysis helps reveal the general trend of a feature's impact on predictions across the dataset. The PD of a feature x_j on the predicted outcome $\hat{y}$ (i.e., log K_d values) is calculated as:

$$\text{PD}_j(x_j) = \frac{1}{n} \sum_{i=1}^{n} f(x_j, x_{i,\setminus j}), \tag{4.9}$$

where $f(x_j, x_{i,\setminus j})$ is the model's prediction when feature x_j is set to a specific value, while all other features $x_{i,\setminus j}$ remain as in each observation i, n is the number of observations, x_j is the feature for which PD is computed, and $x_{i,\setminus j}$ represents the remaining features in each observation, excluding x_j.

4.3 RESULTS AND DISCUSSION

4.3.1 *Effect of chain length in PFAS sorption*

For each entry in the dataset, K_d was converted to K_{OC} to account for variability caused by the differing C_{org} values of the soil (Fabregat-Palau et al., 2021). Then, log K_{OC} was used for assessing the role of PFAS chain length and functional group in sorption. Statistical tests were conducted to group the different log K_{OC} populations (see details in D.5). The mean log K_{OC} data for PFCA ranged from 1.18 to 5.25 and increased with the number of fluorinated carbons, in line with other observations (Figure D.6) (Nguyen et al., 2020; Fabregat-Palau et al., 2021). The K_{OC} populations for PFCA with fewer than six fluorinated carbons were statistically equal and generally increased for each additional CF_2 moiety. Similar outcomes were observed for PFSA (Figure D.7), FOSA (Figure D.8), FTOH (Figure D.9), FTS (Figure D.10), and PFPA (Figure D.11), but this trend was not evident for the cationic and zwitterionic PFAS (Figure D.12) and PFPiA (Figure D.13) subfamilies. The minor dependence of K_{OC} on the chain lengths for short PFAS is evident in our dataset, but it is not entirely clear how potential experimental artifacts may contribute to this (e.g., reliable determination of very small K_d values at large liquid-to-solid ratios and thus potential overestimation). Overall, the log K_{OC} for specific PFAS subfamilies generally increased with increasing chain length, especially for PFAS with six or more fluorinated carbons, likely due to the more effective hydrophobic interaction with soil organic matter (Nguyen et al., 2020; Fabregat-Palau et al., 2021).

4.3.2 *Effect of functional group in PFAS sorption*

Log K_{OC} populations were also assessed to elucidate the effect of the PFAS functional group for those compounds with a certain number of fluorinated carbons (i.e., 4, 6, 8, and 10; see Figures D.14–D.17). Log K_{OC} populations were generally similar regardless of the functional group for those PFAS with less than seven fluorinated carbons (Figures D.14–D.15), but a significant effect of the functional group was noted for PFAS with more than six fluorinated carbons (Figures D.16–D.17). The sorption affinity trend followed: PFPA $\lesssim$ PFCA $\lesssim$ PFSA $\approx$ FTS $<$ FOSA $\lesssim$ FTOH $\lesssim$ cationic and zwitterionic PFAS. The lower sorption observed for PFPA could be attributed to their double negatively charged phosphonate group, in contrast to the single negative charge of the carboxylate group of PFCA, leading to a more effective electrostatic repulsion with the negatively charged surfaces, thereby decreasing its sorption (Du et al., 2014). Higher sorption observed for PFSA compared to that of PFCA agreed with previous observations (Fabregat-Palau et al., 2021) and may result from the greater hydrophobicity provided by the sulfonate moiety. In this line, log K_{OC} values for telomer sulfonate were higher than those for sulfonate groups, likely due to the additional hydrophobicity provided by the $-CH_2-CH_2-$ moiety. Log K_{OC} values for sulfonamide and, especially, telomer alcohol functional groups were significantly higher than others, likely due to their generally uncharged nature, which prevents electrostatic repulsions with negatively charged particles, thereby increasing its sorption (Du et al., 2014). Log K_{OC} values for the cationic and zwitterionic PFAS were the highest, likely a result of the overestimation of K_{OC} due to attraction between the positively charged PFAS species and the negatively charged clay surfaces, thereby increasing sorption (Fabregat-Palau et al., 2024). Similarly, log K_{OC} data for PFAES was the highest observed, likely as a result

of the overestimation of K_{OC} due to data originating from mineral soils (i.e., $C_{org} < 2.7\%$) (Nguyen et al., 2020; Fabregat-Palau et al., 2021).

4.3.3 *Identification of outliers and anomalous trends*

PFAS hydrophobicity (i.e., K_{OW}) increases with the number of fluorinated carbons of the PFAS. According to preliminary data explorations, a positive correlation between K_{OC} and K_{OW} was anticipated due to hydrophobic interaction playing an important role in sorption (Fabregat-Palau et al., 2021). Average log K_{OC} data was therefore screened against log K_{OW}, obtaining a positive linear relationship regardless of the subfamily PFAS type (see Figure 4.2). Nonetheless, four PFAS species (i.e., 6:2 FtSaAm, $C_{6/6}$ PFPiA, $C_{6/8}$ PFPiA, and $C_{8/8}$ PFPiA) exhibited a noted discrepant trend. To the best of our knowledge, sorption of $C_{x/y}$ PFPiA has only been evaluated in a single study (Lee & Mabury, 2017), which concluded that sorption of these species did not increase with organic content, contrary to observations for other PFCA and PFSA (Fabregat-Palau et al., 2021). The obtention of new experimental data supported by additional spectroscopic or computational evidence will be beneficial to confirm if sorption of $C_{x/y}$ PFPiA is driven by other unknown sorption mechanisms. On the other hand, sorption of 6:2 FtSaAm originated from a single soil sample with low organic content (i.e., 0.1%) (Barzen-Hanson et al., 2017), suggesting the overestimation of K_{OC} due to sorption originating mainly from mineral components (Fabregat-Palau et al., 2024). The K_d data for these four PFAS were considered as outliers for further model assessments. The relationship between log K_{OC} and log K_{OW} for PFAS ($\log K_{OC} = 0.58 \log K_{OW} + 0.06; n = 47; r^2 = 0.83$; Figure 4.2) differs from linear correlations observed for other organic pollutants, such as polycyclic aromatic hydrocarbons (PAH, $\log K_{OC} = 0.97 \log K_{OW} + 0.12; n = 106; r^2 = 0.93$) and several non-polar hydrophobic organic compounds (HOC, $\log K_{OC} = 1.10 \log K_{OW} + 0.99; n = 418; r^2 = 0.95$) (Allen-King et al., 2002), likely due to the charged nature of most PFAS considered in this work.

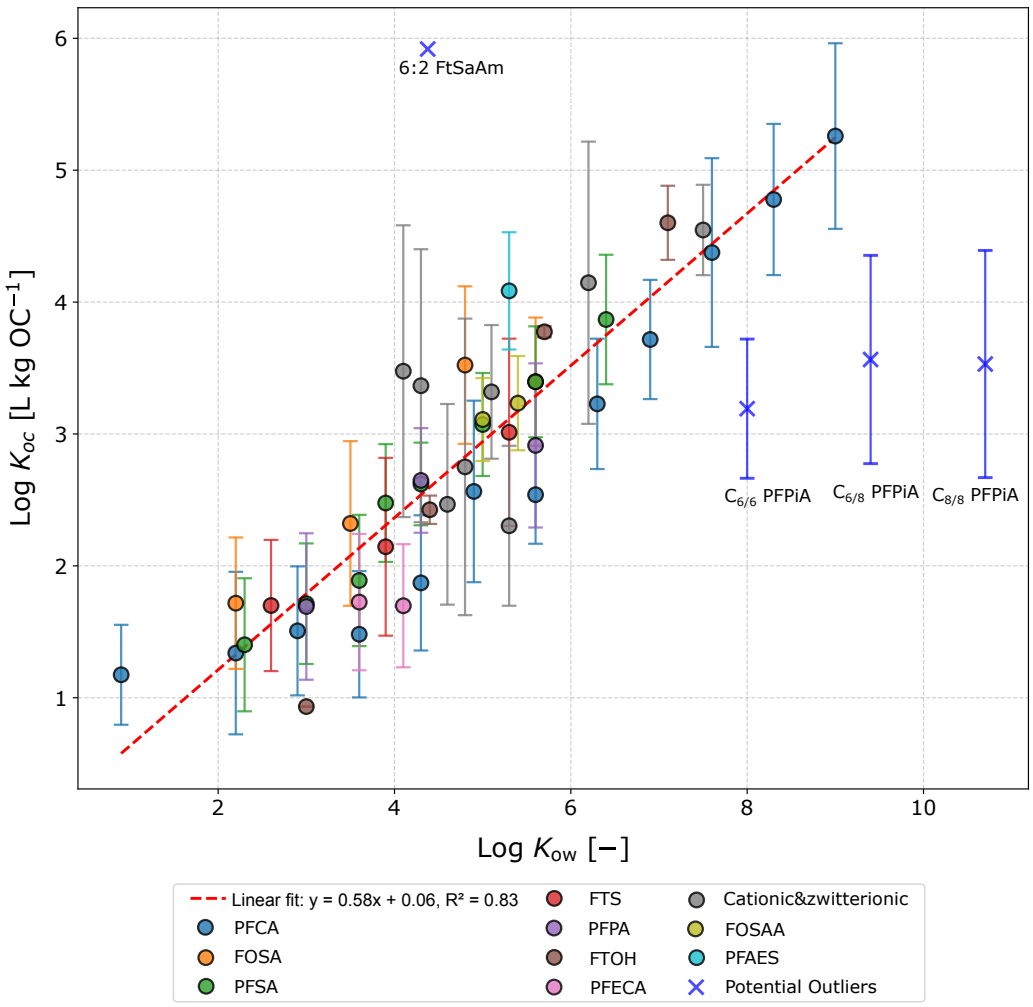


Figure 4.2: Trend between log K_{OC} and log K_{OW} for different PFAS species, including Perfluorocarboxylic Acids (PFCA), Perfluorosulfonamides (FOSA), Perfluorosulfonic Acids (PFSA), Fluorotelomer Sulfonates (FTS), Perfluorophosphonic Acids (PFPA), Fluorotelomer Alcohols (FTOH), Perfluoroalkyl Ether Carboxylic Acids (PFECA), Perfluorooctane Sulfonamidoacetic Acids (FOSAA), and Chlorinated Polyfluoroalkyl Ether Sulfonate (PFAES). Identified PFAS outliers are marked as "X". The dashed line shows a regression of the overall data, excluding outliers. Whiskers represent the standard deviation of the original data.

4.3.4 *Model Performance and Sensitivity Analysis*

For all the PFAS species considered, the predicted log K_d values are in excellent agreement with the observed ones (Figure 4.3), with a slope close to one (i.e., 0.90) and a y-intercept close to zero (i.e., 0.09). NRMSE and RPD values of 0.07 and 3.16, respectively, and a 10 k-fold CV-NRMSE value of 0.09 indicate an excellent prediction ability of the model. Residual evaluations show a residual mean of approximately null log K_d units, evidencing no general under- or over-predictions, and a standard deviation of 0.30. The skewness of the residual distribution was -0.04 log K_d units, with no evident bias along soil C_{org} (Figure 4.3). Additional modeling attempts, based on the same test set, excluding the charge den-

sity feature, resulted in worse predictions of K_d (PFAS), with an NRMSE of 0.08 and an RPD of 2.59.

The PSSM shows better prediction performance compared to other currently available K_d (PFAS) prediction tools, such as those derived from models only considering both organic and mineral sorption sites (Card et al., 2017; Fabregat-Palau et al., 2021) (see D.6). In addition to the good performance for commonly predicted PFCA and PFSA, the model successfully predicts K_d values for TFA in two soils of contrasting characteristics, as well as K_d values for different cationic and zwitterionic PFAS such as PFOSB, PFOAB, PFOAAmS, AmPr-FHxSA, TAmPr-FHxSA, and 6 : 2 FtSaB (Figure 4.3).

Although sorption of ionizable compounds can vary greatly depending on pH due to different speciation (Fabregat-Palau et al., 2024), the accurate prediction for cationic, zwitterionic and FOSA species for soils with differing pH confirms the need for incorporating PFAS speciation into the model. Good predictions were also observed for PFPA, FOSA, FOSAA, PFECA, PFAS, FTS, and FTOH groups. Overall, the successful validation and the high metric qualities of the model highlight its potential use in risk assessment studies aiming to evaluate PFAS mobility in contaminated sites. Nonetheless, while the model demonstrates strong predictive capability for sorption under saturated conditions and effectively incorporates PFAS speciation, its applicability is limited to compounds and soils within the predefined property ranges, as ML struggles to accurately extrapolate predictions to other scenarios. Additionally, our model can be extended by incorporating other physical factors that influence PFAS sorption under unsaturated conditions, such as air-water interfaces, which are ultimately governed by soil moisture and, consequently, subject to seasonal variations. Regarding seasonal variations, the model so far does not account for temperature-induced changes in K_d values for PFAS (according to Xiang et al., 2018, K_d may increase by up to 60% if temperatures rise from 15 to 35$^{\circ}C$) .

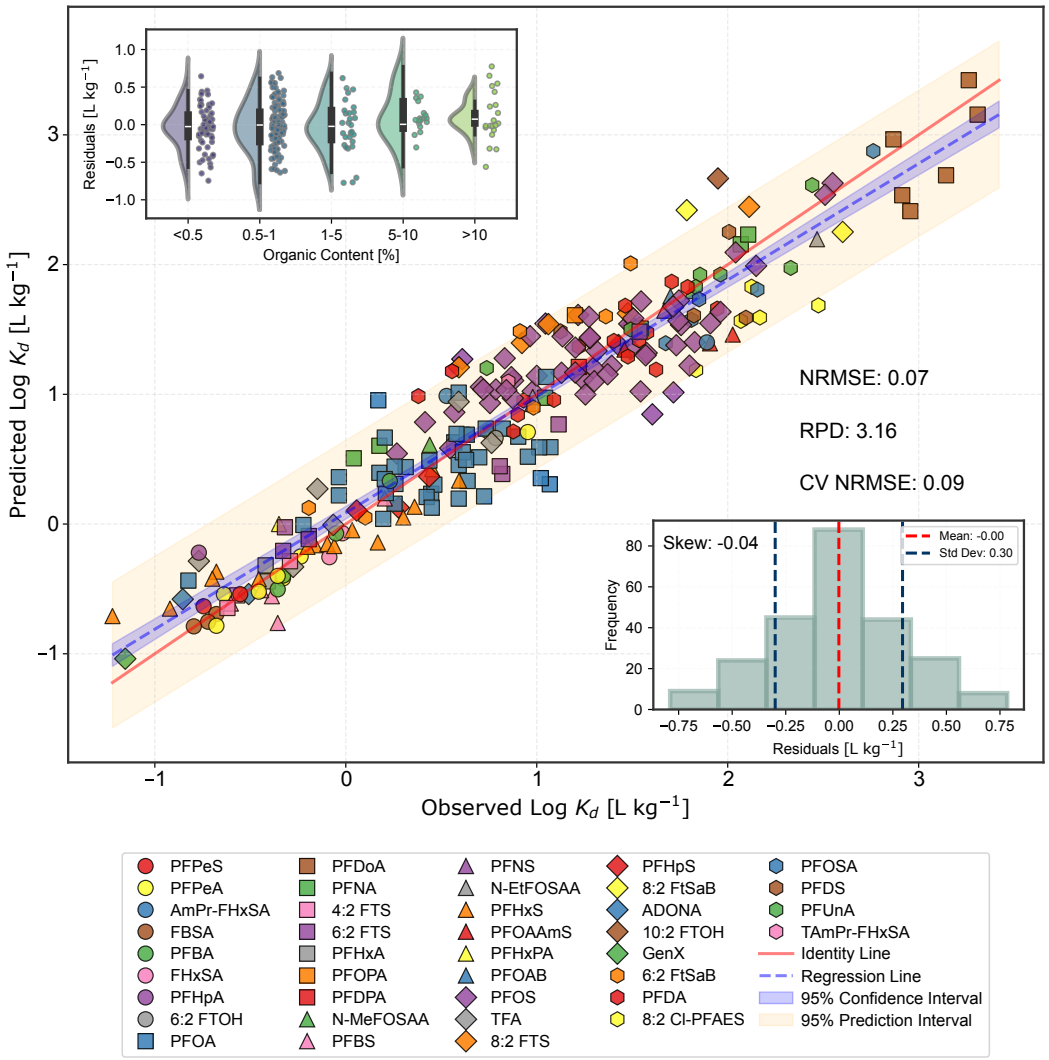


Figure 4.3: Scatter plot comparing predicted and measured log K_d values for various PFAS species. The lower-right subplot illustrates the distribution of residuals, with a mean of approximately zero log K_d units, a standard deviation of 0.30, and a skewness of −0.04 log K_d units. The upper-left subplot represents the residual distribution stratified by different levels of soil organic content.

The sensitivity analysis (Figure 4.4) reveals that MW is the most sensitive feature, particularly distinguishing PFAS species. Higher MW values correlate with increased log K_d, as shown by both SHAP values (Figure 4.4A) and PD analysis (Figure 4.4B). Unlike the findings of Xie and coworkers (Xie et al., 2024), who identified pH as the second most influential feature, our analysis shows C_{org} as the predominant soil property influencing sorption. This result aligns with the strong affinity of PFAS for organic matter, indicating that the key influence of pH in Xie's model may originate from including K_d values for the same PFAS/soil combinations at different pH levels. In our model, C_{org} contributes positively to log K_d, suggesting organic carbon content as a key soil driver of PFAS sorption, as confirmed by both SHAP and PD analyses. Log K_{OW} emerged as the third most important feature, with both SHAP and PD indicating that high log K_{OW} values correlate with high log K_d (also supported by findings in Figure 4.2). The influence of both C_{org} and K_{OW} supports the hypothesis that hydrophobic interactions are a central mechanism

for PFAS sorption (Fabregat-Palau et al., 2021). Additionally, charge density was identified as the fourth most influential feature, suggesting minimal impact of electrostatic interaction on PFAS sorption. CEC ranked fifth in influence, with sorption increasing alongside higher CEC values, potentially due to indirect associations with higher C_{org} contents and increased availability of exchangeable sorption sites, especially suitable to sorb cationic and zwitterionic species (Barzen-Hanson et al., 2017). Conversely, increasing pH generally decreased K_d (PFAS), likely due to promoting a higher abundance of negatively charged PFAS species and an increasing number of negative charges in clay ($pH_{ZPC} \approx 3$) and C_{org} ($pH_{ZPC} \approx 8$) fractions (Fabregat-Palau et al., 2024), thus increasing electrostatic repulsions with these domains. Finally, soil texture had a limited effect on the model's predictions, suggesting that grain sizes play a minor role in PFAS sorption compared to C_{org}. However, including these variables may still enhance prediction accuracy in low C_{org} scenarios (Fabregat-Palau et al., 2021).

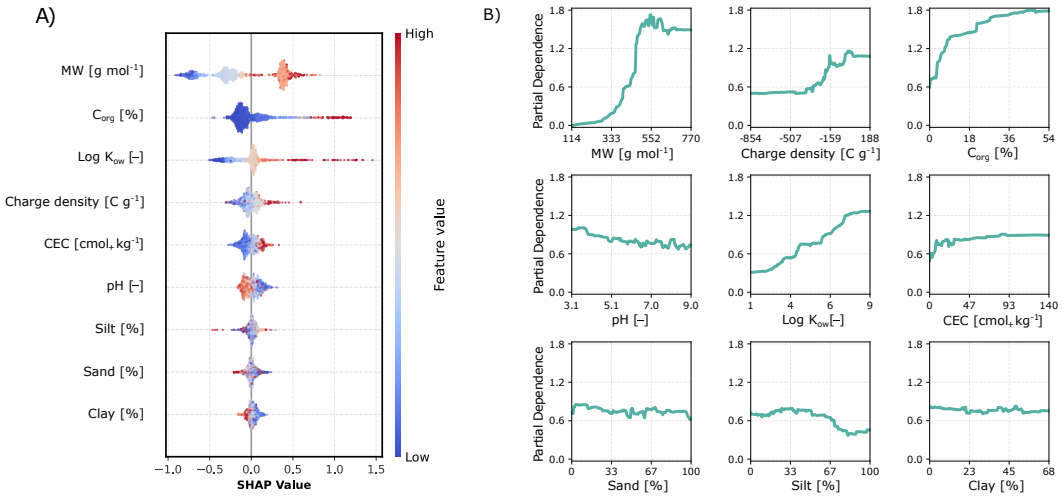


Figure 4.4: A) SHAP summary plot illustrating the influence of individual features on K_d predictions. Features are ranked from top to bottom by their importance on model's output, with each dot representing a single sample. The color gradient from blue (low) to red (high) indicates the feature value, while the horizontal position reflects the model's output change, where negative SHAP values correspond to lower output values and positive SHAP values indicate higher output values. (B) Partial Dependence (PD) showing the marginal effect of each input feature on K_d, averaged across all samples.

4.4 ENVIRONMENTAL IMPLICATIONS

The developed model (i.e., PSSM) offers broad applicability to predict K_d for 47 PFAS by leveraging both PFAS-specific and soil-dependent input parameters. With this flexibility, the model can be applied to spatial soil property data repositories at any scale and resolution to generate K_d (PFAS) prediction maps. To give an example, we applied the PSSM to the soil LUCAS 2009 repository data for European soils (European Soil Data Centre: ESDAC). Figure 4.5 shows the K_d map for PFOSB, while additional information and K_d maps for other PFAS species (i.e., TFA, PFOA, PFOS) are available in D.7. Overall, the model allows the identification of geographical zones with potentially low K_d values (i.e., locations where PFAS may have higher mobility), thus facilitating its transport to the groundwater table. Therefore, the combination of our model output with other PFAS topsoil concentra-

tion data and groundwater information can be used in global PFAS transport studies across the saturated zone (Guo et al., 2020; Moghadasi et al., 2023). However, these large-scale assessments may overlook local heterogeneities in soil properties, potentially introducing additional uncertainty in K_d predictions unless high-resolution geospatial data is available.

To overcome these soil spatial heterogeneity limitations, we provide an online platform (PFASorptionML) for end-users to predict site-specific K_d (PFAS) values. The platform is free to use at https://hydrogeochem.geo.uni-tuebingen.de/pfas. Input parameters required are the PFAS to be assessed (i.e., CAS number, full name or abbreviation of the PFAS) and specific soil properties (i.e., pH, C_{org}, CEC, soil texture). In case users lack some of these soil properties, our KNN imputer model developed in D.4 facilitates their prediction. Using this information, PFAS-specific pH-speciation plots are generated and K_d predictions are produced. Furthermore, the platform allows the generation of K_d maps at any scale (e.g., EU) for all 47 PFAS considered in this study. Further assessments of the model toward other PFAS that currently lack sorption data are of interest and subject to further work. Upgrades of the platform are planned, including the generation of additional geographical K_d maps (e.g., Worldwide) and the consideration of additional transport parameters such as specific air-water interfaces and K_{aw} in unsaturated conditions. In addition to topsoil properties, sub-soils can be included to allow more realistic transport modeling across the vadose zone into groundwater. The findings of this study, considering the limitations of the PSSM model, can be integrated into risk assessment evaluations to assess the mobility of PFAS in soils and aquifers.

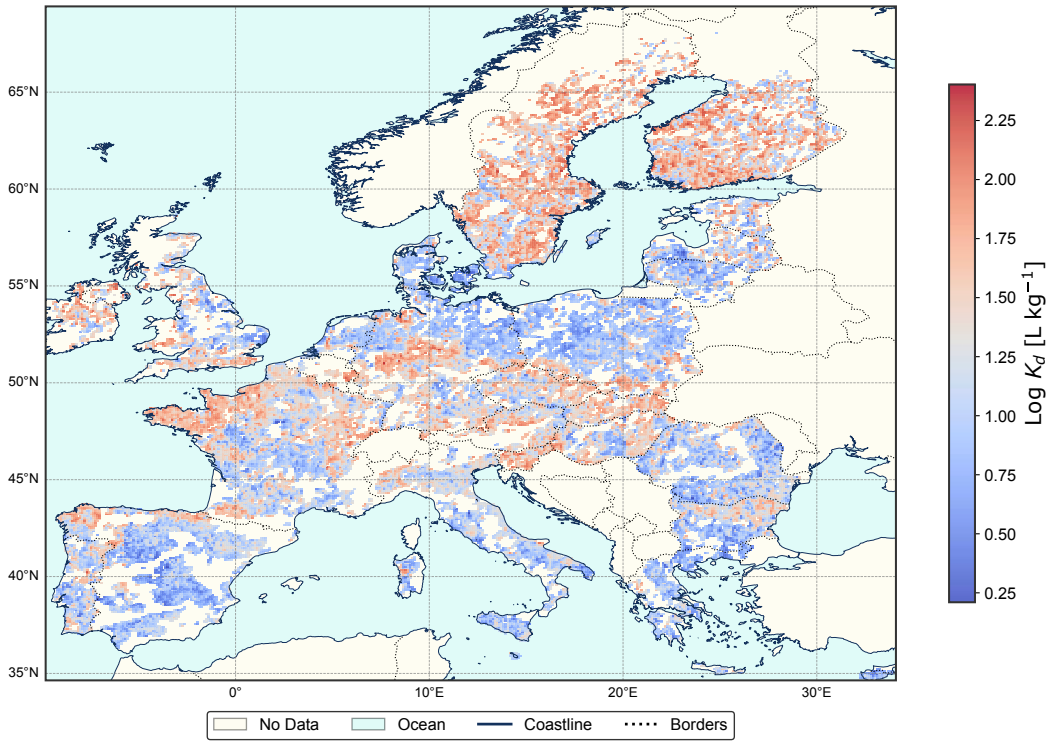


Figure 4.5: Predicted log K_d values for PFOSB across Europe using the stacking model based on soil properties from the LUCAS 2009 repository. Regions without available soil data are indicated as "No Data".

5

CONCLUSION&OUTLOOK

5.1 SYNTHESIS OF MAJOR FINDINGS

This thesis addresses challenges in contaminant assessment by introducing efficient and scalable, data-driven methodologies designed to predict the leaching and sorption behavior of various environmental pollutants. By integrating ML models with comprehensive empirical datasets, the research significantly reduces reliance on traditional, time-intensive laboratory experiments and computationally demanding simulations. These innovative approaches not only streamline the prediction process but also enhance overall efficiency while maintaining high accuracy. Furthermore, the scalability of these solutions ensures their applicability across a wide range of contaminants and environmental conditions, making them valuable tools for researchers and practitioners in environmental science and engineering.

5.1.1 *How does machine learning transform leaching test assessments?*

The application of ML represents a paradigm shift in leaching test assessments, enabling rapid and accurate predictions of leachate concentrations across extended liquid-to-solid ratios. By leveraging readily available parameters—such as pH, electrical conductivity, and compound-specific leachate concentration—ML models provide results that are computationally efficient and robust. This study demonstrates the suitability of ML models for predicting long-term leaching behavior based on short-term concentration data (LS$\leq$1), significantly reducing laboratory testing timelines from weeks to hours. The models accurately forecasted long-term concentrations (LS=2, 4, 10) and produced cumulative concentration predictions that aligned well with measured values. This approach presents a viable alternative for regulatory compliance and material categorization under German recycling guidelines (DIN 19528, 2009), ensuring that classification accuracy is maintained while substantially expediting the assessment process. Beyond accelerating assessments, these models, using SHAP values for sensitivity analysis, offer insights into the key factors governing leaching dynamics, such as pH. By enabling faster, data-driven decision-making, these models help ensure the safe reuse of materials while supporting environmental protection efforts, contributing to broader sustainability goals.

5.1.2 *How are computational bottlenecks in leaching simulations overcome?*

This thesis introduces ensemble surrogate models that emulate one-dimensional advective-dispersive transport coupled with intraparticle diffusion, a process traditionally simulated using a cell-centered finite volume method. These surrogates replicate the underlying physics of contaminant leaching while achieving a 1,000-fold computational speedup, reducing simulation times from over 150 seconds to less than 0.1 seconds. By leveraging adaptive-recursive sampling strategies that balance exploration (via uncertainty assessment) and exploitation (via error minimization), the models achieve high accuracy across

the parameter space defined in DIN 19528, 2009, with a median RRMSE of 0.11. The framework's robustness is further enhanced by integrating Simulation-Based Inference (SBI), which eliminates the need for explicit likelihood evaluations—a key limitation of traditional Bayesian methods in high-dimensional settings. By mapping parameters directly to observations via surrogate models, SBI enables efficient posterior inference, as demonstrated by neural posterior estimation applied to surrogate-generated simulations. This approach accurately reconstructed cumulative copper concentration measurements from leaching tests on two distinct soil types, underscoring its effectiveness in capturing complex contaminant transport behavior while ensuring robust and computationally efficient parameter estimation.

5.1.3 *What new insights are provided into PFAS sorption behavior?*

The PFASorptionML tool developed in this thesis offers a groundbreaking framework for understanding and predicting PFAS sorption in soils. By integrating compound properties such as molecular weight and hydrophobicity (i.e., log K_{ow}) with soil-specific factors like organic carbon content, pH, and soil texture, the ML-based stacking model accurately predicts the solid-liquid distribution coefficient (K_d) for a diverse range of PFAS in soils, while revealing the key factors that influence their mobility through SHAP values and partial dependence analysis. The generation of high-resolution EU K_d maps, created using soil data from the LUCAS soil database, highlights the practical applications of this tool in assessing PFAS risks across diverse environmental contexts. These maps highlight regions with varying K_d values, identifying areas where PFAS mobility to groundwater may be elevated. The accompanying open-access platform (https://hydrogeochem.geo.uni-tuebingen.de/pfas) enables stakeholders to predict site-specific K_d values for 47 PFAS species, even with incomplete soil data through an integrated KNN imputer.

5.2 FUTURE PROSPECTS

5.2.1 *Expanding Datasets for Enhanced Predictive Performance*

The expansion of datasets is essential to improving the generalizability, accuracy, and reliability of the methodologies developed in this thesis. For column leaching tests, incorporating data from multiple regulatory frameworks—such as the *EU CEN/TS standards* and the *US Environmental Protection Agency guidelines*—will ensure that models remain adaptable to varying compliance landscapes. Including additional material types, such as contaminated soils, mining residues, and industrial byproducts, will refine predictive accuracy and broaden the model's applicability across diverse environmental contexts.

For PFAS sorption assessments, expanding datasets to cover a more extensive range of PFAS classes—including Betaine compounds with pH-dependent speciation—will significantly enhance predictive performance. Additionally, integrating geochemical parameters (e.g., Fe and Al content), mineralogical data, and porosity characteristics will improve model precision in capturing complex sorption interactions across diverse environmental conditions. A more comprehensive dataset with denser measurements across varying pH levels, organic carbon content, and soil texture compositions will be essential for refining model accuracy.

5.2.2 *Broadening the Scope to Emerging Contaminants*

The methodologies developed in this thesis offer a scalable foundation for evaluating a wider spectrum of emerging contaminants, including antibiotics, microplastics, heavy metals, and other persistent pollutants. These contaminants exhibit distinct chemical behaviors and environmental interactions that require refined predictive modeling approaches. By incorporating targeted datasets and adapting model parameters accordingly, the predictive capabilities established in this work can be extended to address these emerging challenges.

For instance, antibiotic contamination poses a growing environmental risk due to its potential role in fostering antimicrobial resistance. Similarly, microplastics, characterized by their size-dependent transport dynamics and sorption behavior, require specialized modeling approaches to assess their mobility in terrestrial and aquatic systems. Expanding models to incorporate these contaminants will enhance their applicability in environmental risk assessment and regulatory decision-making.

5.2.3 *Applying Sorption Models on a Global Scale*

To ensure broad applicability, expanding datasets to include contamination profiles from underrepresented regions is crucial. Environmental factors such as diverse soil compositions, organic carbon content, and pH variations must be comprehensively represented to develop risk assessment models that remain relevant across diverse geochemical settings.

Developing global-scale sorption models will facilitate high-resolution global sorption maps, aiding in the identification of contamination hotspots and supporting large-scale environmental assessments. In particular, including contamination data from developing regions—where environmental conditions and pollution profiles often differ significantly from well-studied locations—will strengthen model applicability. By integrating globally

representative datasets, these models will improve international risk assessments and inform coordinated remediation efforts, contributing to global environmental policy development.

5.2.4 *Application of Sorption Coefficients in Contaminant Transport Models*

A key advancement in this work is the integration of sorption coefficients (K_d) into contaminant transport models. Predicting the mobility of contaminants in both saturated and unsaturated zones using these values will improve simulations of groundwater transport, facilitating accurate predictions of contaminant fate under varying environmental conditions. For instance, these models can estimate the travel time of contaminants to groundwater, offering crucial insights into potential risks and transport behavior. Such information is essential for groundwater protection, site-specific risk assessments, and the development of effective remediation strategies.

5.2.5 *Extending Surrogate Models for Broader Environmental Applications*

The surrogate models developed in this thesis, designed to replicate one-dimensional advection, dispersion coupled with intraparticle pore diffusion, were primarily trained on datasets from German standard column leaching tests. These models offer a computationally efficient alternative to traditional contaminant leaching assessments. However, their applicability can be expanded by integrating additional regulatory protocols, such as stop-flow tests and international leaching standards from the European Union and the United States.

Further advancements in surrogate modeling should focus on alternative material compositions (e.g., industrial byproducts and construction waste) to improve predictive capabilities. Additionally, hybrid modeling approaches—such as combining machine learning with mechanistic simulations like physics-informed neural networks—will enhance the models' ability to capture complex contaminant transport processes.

5.2.6 *Towards Integrated Environmental Management Platforms*

Future developments should focus on integrating the methodologies established in this thesis into a comprehensive environmental risk management platform. By coupling sorption predictions, leaching models, and transport simulations into a unified system, such a platform could provide real-time insights into contaminant fate and mobility, supporting evidence-based decision-making.

Incorporating interactive visualization tools, real-time data integration, and user-friendly interfaces will enhance accessibility for researchers, regulatory agencies, and industry stakeholders. This integrated approach will facilitate large-scale environmental assessments, enable site-specific risk evaluations, and contribute to more effective contamination mitigation strategies.

A

APPENDIX FOR CHAPTER 1

A.1 OVERVIEW OF 15 PRIORITY PAHs BY US-EPA

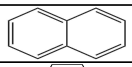
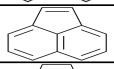
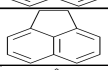
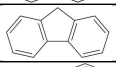
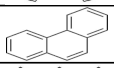
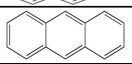
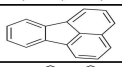
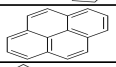
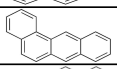
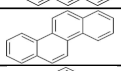
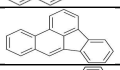
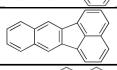
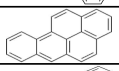
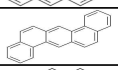
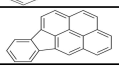
Compound Name	Molecular Formula	Compound Structure
Naphthalene	$C_{10}H_8$	
Acenaphthylene	$C_{12}H_8$	
Acenaphthene	$C_{12}H_{10}$	
Fluorene	$C_{13}H_{10}$	
Phenanthrene	$C_{14}H_{10}$	
Anthracene	$C_{14}H_{10}$	
Fluoranthene	$C_{16}H_{10}$	
Pyrene	$C_{16}H_{10}$	
Benzo[a]anthracene	$C_{18}H_{12}$	
Chrysene	$C_{18}H_{12}$	
Benzo[b]fluoranthene	$C_{20}H_{12}$	
Benzo[k]fluoranthene	$C_{20}H_{12}$	
Benzo[a]pyrene	$C_{20}H_{12}$	
Dibenzo[a,h]anthracene	$C_{22}H_{14}$	
Indeno[1,2,3,cd]pyrene	$C_{22}H_{12}$	

Table A.1: The 15 (US-EPA) PAHs along with their molecular formulas and structures

A.2 COLUMN LEACHING TEST STANDARDS

Parameter	Unit	DIN 19528	EPA 1314	EN 14405	CEN/TS 16637-3	ISO/TS 21268-3
Column diameter (D)	cm	5–10	5	5-10	5-10	5-10
Column height (H)	cm	$> 4 \cdot D$	$\approx 28 \pm 3$	30 ± 5	30 ± 5	30 ± 5
Grain size	mm	≤ 32 and $\leq 0.5 \cdot D$	≤ 2.5	$\leq 0.1 \cdot D$[a]	≤ 4	≤ 4
Flow rate	cm d^{-1}	40–80	33 ± 12[b]	30 ± 4	15 ± 2	30 ± 4
Eluent	-	deionized water	deionized water	deionized water	deionized water	0.001 mol L^{-1} $CaCl_2$
Saturation time	h	2	$6 - 12$	not specified	4 ± 2[b]	not specified
Pre-equilibration	h	0	21 ± 3	16-72	12-72	≥ 48
Contact time during percolation	h	5	6–12[2]	~ 20	~ 10	~ 20
Temperature	°C	20 ± 2	not specified	22 ± 3	22 ± 3	20 ± 5
Eluate collection (LS)	L kg^{-1}	0.3, 0.5, 1.0, 2.0, 5.0, 10.0	0.2, 0.5, 1.0, 1.5, 2.0, 4.5, 5.0, 9.5, 10.0	0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0	0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0	0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0
Sand admixture allowed	-	yes	no	no	no	no

[a] Requirement for large columns (diameter 10 cm). For smaller columns (diameter 5 cm): 95 wt.% (weight percent) of particles must be <4 mm (without size reduction) or 80–95 wt.% <4 mm and up to 5 wt.% <10 mm (with size reduction of the <4 mm fraction). Weight percent (wt.%) represents the proportion of a substance's weight relative to the total weight of the mixture.
[b] Not specified. The provided values were calculated as examples for a material with a porosity of 0.4 and a packing density of 1.6 g cm^{-3} (Lin et al., 2020).

Table A.2: Overview of column leaching test standards

A.3 MISSING DATA IMPUTATION

A.3.1 *Basic imputation techniques*

Mean and median imputation are foundational methods for handling missing data. These approaches replace missing values with the mean or median of the observed data within the same feature. In mean imputation, the missing value x_i is replaced by the mean of the observed values x_j, given as:

$$x_i = \frac{1}{n} \sum_{j=1}^{n} x_j, \tag{A.1}$$

where n denotes the total number of observed values in the feature.

For median imputation, the missing value x_i is replaced by the median of the observed values. Let the observed values $x_1, x_2, \ldots, x_n$ be sorted in ascending order. The imputation formula is:

$$x_i = \begin{cases} x_{(m)}, & \text{if } n \text{ is odd} \\ \frac{x_{(m)} + x_{(m+1)}}{2}, & \text{if } n \text{ is even} \end{cases} \tag{A.2}$$

where $x_{(m)}$ is the middle value in the sorted sequence. These methods are computationally efficient and easy to implement but may distort the underlying data distribution and fail to preserve relationships among variables, leading to biased analyses.

A.3.2 *k-nearest neighbors (KNN)*

The KNN imputer identifies the k most similar instances (neighbors) to a data point with missing values (Troyanskaya et al., 2001). It computes the similarity between data points based on the distance in feature space, typically using Euclidean distance. For each missing value, the imputer finds the k−nearest data points with known values for that feature and imputes the missing value as the weighted average (or median) of the neighbors' values. The Euclidean distance $d(x_i, x_j)$ between two data points x_i, x_j, each represented by a set of features, is calculated as Equation A.3:

$$d\left(x_i, x_j\right) = \sqrt{\sum_{f=1}^{n} \left(x_{if} - x_{jf}\right)^2}, \tag{A.3}$$

where x_i, x_j are the values of the feature f for the data points x_i, x_j, and n is the total number of features. Once these distances are calculated, the KNN are selected. The imputed value for the missing feature v_{missing} is then estimated as Equation A.4:

$$v_{\text{missing}} = \frac{1}{k} \sum_{i=1}^{k} v_i, \tag{A.4}$$

where v_i represents the values of the feature from the KNN. This approach ensures that imputed values reflect the patterns and relationships present in the surrounding data, leveraging the inherent spatial and feature-based correlations in soil properties. By employing KNN imputation, we preserved the structure and variability of the original dataset while minimizing bias that could arise from removing incomplete data (Troyanskaya et al., 2001).

A.4 DATA SCALING

Standardization, also known as Z-score scaling, transforms data to have a mean of zero and a standard deviation of one. For a given feature x, the standardized value z is computed as:

$$z = \frac{x - \mu}{\sigma}, \tag{A.5}$$

where x is the original feature value, μ is the mean of the feature, and σ is its standard deviation. This transformation centers the data around zero and scales it to unit variance, ensuring that features with larger magnitudes do not dominate the learning process.

Min-Max scaling is another commonly used method that normalizes feature values to a fixed range, typically between [0, 1]. The Min-Max scaled value x' for a feature x is calculated as:

$$x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}, \tag{A.6}$$

where x_{min} and x_{max} represent the minimum and maximum values of the feature, respectively. This method preserves the relative distribution of the data while making all features comparable in magnitude. Min-Max scaling is particularly beneficial for algorithms requiring bounded input values, such as neural networks, where large feature values may hinder convergence.

A.5 UNSUPERVISED AND REINFORCEMENT LEARNING

Unsupervised learning focuses on analyzing datasets without labels, uncovering hidden structures and relationships (Ghahramani, 2004). Techniques like clustering, dimensionality reduction, and anomaly detection enable exploratory analysis and data preprocessing. Clustering algorithms, such as *k-means* and *Gaussian Mixture Models*, group data points based on similarity, with applications in identifying patterns within datasets (Yuan et al., 2014; Murtagh & Contreras, 2017; Sinaga & Yang, 2020; Li et al., 2021). Dimensionality reduction techniques, including *Principal Component Analysis* and *t-distributed Stochastic Neighbor Embedding*, simplify high-dimensional data while preserving its variability, enhancing computational efficiency and visualization (Ma & Yuan, 2019; Belkina et al., 2019). Anomaly detection methods like *Isolation Forest* and *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)* identify irregularities in large and noisy datasets, making them valuable for outlier detection (Thang & Kim, 2011).

Reinforcement learning (RL), on the other hand, addresses sequential decision-making problems by enabling agents to interact with an environment and learn from feedback (Russell & Norvig, 2022). Unlike unsupervised learning, RL relies on trial and error to optimize cumulative rewards. RL is particularly suited for dynamic and adaptive systems, with applications in robotics, enabling autonomous navigation, and resource management.

A.6 LINEAR REGRESSION

Linear regression models the relationship between input features X and the target output Y using the equation:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon, \tag{A.7}$$

where β_0 is the intercept, β_i are the coefficients, and ε is the residual error. The coefficients β_i are estimated by minimizing the residual sum of squares (RSS):

$$L_{\text{Linear}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{A.8}$$

To address overfitting, Ridge regression introduces an L_2-norm penalty:

$$L_{\text{Ridge}} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{n} \beta_j^2, \tag{A.9}$$

where $\lambda > 0$ controls the strength of regularization, discouraging large coefficients.

Lasso regression incorporates an L_1-norm penalty, which promotes sparsity:

$$L_{\text{Lasso}} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{n} |\beta_j|. \tag{A.10}$$

This penalty reduces some coefficients to exactly zero, enabling feature selection. These regularization techniques enhance generalization and reduce model complexity.

A.7 ENSEMBLE ALGORITHMS

This section details the mathematical formulations and mechanisms of key ensemble methods employed to enhance predictive accuracy and robustness. Each method leverages distinct strategies to address the limitations of individual models.

A.7.1 *Random Forest*

Random Forest constructs multiple decision trees using bootstrapped subsets of the training data (Breiman, 2001). At each node, it selects a random subset of features to determine the optimal split, reducing overfitting and enhancing model generalization. For regression tasks, the final prediction is computed as the average of the predictions from all trees:

$$\hat{Y} = \frac{1}{T} \sum_{t=1}^{T} \hat{f}_t(X), \tag{A.11}$$

where T is the total number of trees, $\hat{f}_t(X)$ is the prediction of the t-th tree, and $\hat{Y}$ is the final aggregated prediction. Each decision tree is constructed by recursively partitioning the input space to minimize the variance in the target variable. The optimal split at each node is determined by maximizing the reduction in impurity, often measured by the mean squared error (MSE) in regression (see Equation A.17).

A.7.2 *Extremely Randomized Trees (ExtraTrees)*

Extremely Randomized Trees (ExtraTrees) is a tree-based ensemble method designed to enhance predictive diversity and robustness by incorporating additional randomness into both data sampling and feature splitting processes (Geurts et al., 2006). While similar to Random Forest, ExtraTrees introduces key differences in its approach to tree construction.

Unlike Random Forest, where split points are chosen to minimize a specific loss function, ExtraTrees selects split points randomly within the range of the selected feature. This process bypasses the optimization step typically performed during the construction of individual trees, significantly reducing computational overhead. The randomness introduced during split selection leads to increased variance among the trees in the ensemble, making the model more robust to noise and reducing overfitting in high-dimensional datasets or datasets with noisy features.

For each tree in the ensemble: 1. A random subset of features is selected at each node. 2. Split thresholds for these features are determined randomly, rather than being optimized based on the loss function. This dual-randomization strategy results in highly diverse individual trees, which, when aggregated, yield a model with improved generalization capabilities. For regression tasks, the final prediction is computed as the average of predictions from all trees in the ensemble (see Equation A.11).

A.7.3 *Gradient Boosting*

Gradient Boosting is an ensemble learning technique that builds models sequentially, with each subsequent model correcting the prediction errors of the combined model from pre-

vious iterations (Friedman, 2001). The core idea is to minimize a specified loss function by fitting a new model to the negative gradient (pseudo-residuals) of the loss at each iteration.

At iteration m, the model prediction is updated as:

$$\hat{Y}^{(m)} = \hat{Y}^{(m-1)} + \eta \cdot h_m(X), \tag{A.12}$$

where $\hat{Y}^{(m)}$ is the updated prediction, $\hat{Y}^{(m-1)}$ is the prediction from the previous iteration, η is the learning rate, which controls the contribution of each weak learner, and $h_m(X)$ is the weak learner trained on the pseudo-residuals:

$$r_i^{(m)} = -\frac{\partial \mathcal{L}(Y_i, \hat{Y}_i^{(m-1)})}{\partial \hat{Y}_i^{(m-1)}}, \tag{A.13}$$

with $r_i^{(m)}$ representing the pseudo-residuals, Y_i is the true target value, $\hat{Y}_i^{(m-1)}$ is the predicted value from the previous iteration, and $\mathcal{L}$ is the chosen loss function.

A.7.4 *Extreme Gradient Boosting (XGBoost)*

XGBoost optimizes Gradient Boosting by incorporating regularization, parallelization, and sparsity-aware computations (Chen & Guestrin, 2016). Its objective function combines the loss function and a regularization term:

$$L = \sum_{i=1}^{N} \mathcal{L}(Y_i, \hat{Y}_i) + \sum_{k=1}^{M} \Omega(f_k), \tag{A.14}$$

where N is the number of training samples, $\mathcal{L}(Y_i, \hat{Y}_i)$ measures the difference between the observed target Y_i and the predicted value $\hat{Y}_i$, M is the number of trees in the model, and $\Omega(f_k)$ penalizes the complexity of the k-th tree:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{A.15}$$

where T is the number of leaf nodes, w_j are the leaf weights, γ controls the number of leaves, and λ applies L2 regularization.

Each subsequent tree minimizes the residuals, updating the model iteratively:

$$\hat{Y}^{(t)} = \hat{Y}^{(t-1)} + \eta f_t(X), \tag{A.16}$$

where $f_t(X)$ is the t-th tree.

A.8 TRAINING PROCESS OF MULTI-LAYER PERCEPTRONS (MLP)

The training of Multi-Layer Perceptrons (MLPs) involves several steps to iteratively adjust parameters (weights and biases) for improving predictive performance.

A.8.1 *Core training steps*

FORWARD PROPAGATION: During forward propagation, input data X is sequentially passed through each layer of the MLP. Each layer applies a linear transformation followed by a non-linear activation function. Common activation functions include:

- **ReLU (Rectified Linear Unit):** Defined as $\sigma(x) = \max(0, x)$, where x is the input to a neuron. ReLU is widely used due to its simplicity and ability to mitigate the vanishing gradient problem.
- **Sigmoid:** Defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, it maps inputs to a range of $(0, 1)$ and is commonly used for binary classification tasks.
- **Tanh (Hyperbolic Tangent):** Defined as $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$, it maps inputs to a range of $(-1, 1)$, making it suitable for zero-centered data.

The forward propagation process continues through hidden layers and concludes with the output layer, where predictions $\hat{Y}$ are generated.

LOSS CALCULATION: The loss function L quantifies the difference between the predicted outputs $\hat{Y}$ and the true values Y, providing an objective measure of the model's performance and guiding the optimization process.

BACKWARD PROPAGATION: Backward propagation computes the gradients of the loss function L with respect to the model parameters (weights and biases). Using the chain rule of differentiation, gradients are calculated layer by layer, starting from the output layer and propagating backward to the input layer. These gradients guide the parameter updates.

WEIGHT UPDATE: Weights and biases are updated using an optimization algorithm such as Stochastic Gradient Descent (SGD) or Adam. The update rule for a parameter θ is given by:

$$\theta \leftarrow \theta - \eta \frac{\partial L}{\partial \theta},$$

where η is the learning rate, controlling the step size for each update.

A.8.2 *Regularization techniques*

Regularization helps prevent overfitting by introducing constraints during training. Common techniques include:

DROPOUT: Dropout randomly sets a fraction of neuron activations to zero during training. This reduces reliance on specific neurons and enhances model generalization.

WEIGHT DECAY: Weight decay adds a penalty term to the loss function, encouraging smaller weight magnitudes and reducing overfitting. The regularized loss is given by:

$$L_{\text{reg}} = L + \lambda \sum_{j} w_j^2,$$

where λ is the regularization strength, and w_j are the weights.

A.9 MODEL PERFORMANCE EVALUATION METRICS

Quantitative evaluation of model performance is essential for assessing the accuracy and reliability of predictions. Several commonly used metrics are described below:

The *Mean Squared Error (MSE)* quantifies the average squared difference between observed and predicted values:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2, \tag{A.17}$$

where Y_i and $\hat{Y}_i$ are the observed and predicted values, respectively, and N is the total number of data points. A lower MSE indicates better model performance, as it reflects smaller deviations between predictions and observations.

The *Mean Absolute Error (MAE)* represents the average absolute difference between observed and predicted values:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right|, \tag{A.18}$$

providing a more interpretable metric in the original scale of the data. Like MSE, a lower MAE signifies improved prediction accuracy, though it is less sensitive to large errors.

The *Root Mean Squared Error (RMSE)* represents the square root of the MSE, offering an interpretable measure of error magnitude in the same units as the observed data:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}. \tag{A.19}$$

RMSE is particularly sensitive to large errors, making it suitable for identifying significant prediction discrepancies.

The *Relative Root Mean Squared Error (RRMSE)* evaluates the relative error magnitude by normalizing the RMSE by the mean of observed values:

$$\mathrm{RRMSE} = \frac{\mathrm{RMSE}}{\overline{Y}} \times 100, \tag{A.20}$$

where $\overline{Y}$ is the mean of the observed values. RRMSE is expressed as a percentage, making it useful for comparing error magnitudes across datasets with different scales.

The *Determination Coefficient (R^2)* measures the proportion of variance in the observed data explained by the model:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}. \tag{A.21}$$

An R^2 value close to 1 indicates that the model explains most of the variance in the data, whereas values closer to 0 suggest poor predictive capability.

The *Ratio of Performance to Deviation (RPD)* assesses the predictive accuracy of the model relative to the variability of the dataset:

$$\mathrm{RPD} = \frac{\sigma_Y}{\mathrm{RMSE}}, \tag{A.22}$$

where σ_Y is the standard deviation of the observed values. For environmental analysis, RPD < 1.5 values indicate poor performance, while RPD values ranging from 1.5 to 2.0 indicate acceptable quality. RPD values > 2.0 indicate good quality, especially those with RPD ≥ 3.0, which are considered as analytical quality (Knight et al., 2019)

The *K-fold Cross-validated Normalized Root Mean Square Error (CV-NRMSE)* is a metric used to assess the model's performance across different folds during cross-validation. By normalizing the RMSE based on the data's range, this metric provides an estimate of how well the model generalizes to unseen data. It is calculated similarly to NRMSE but is specifically applied during the cross-validation process:

$$\text{CV-NRMSE} = \frac{\sqrt{\frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\left(y_i^{(j)} - \hat{y_i^{(j)}}\right)^2\right)}}{y_{\max} - y_{\min}}, \tag{A.23}$$

where k is the number of cross-validation folds, n_j is the number of samples in fold j, and $y_i^{(j)}$ and $\hat{y_i^{(j)}}$ are the measured and predicted values for fold j. The CV-NRMSE helps ensure that the model's error estimates are reliable and not overfitted to a particular training set.

B

APPENDIX FOR CHAPTER 2

B.1 RELEVANT COMPOUNDS AND THE THRESHOLD CONCENTRATION IN AQUEOUS LEACHATE ACCORDING TO GERMAN RECYCLING DECREE

Model		CDW-1	CDW-2	CDW-3
Sulfate	[mg L^{-1}]	600	1000	3500
Vanadium	[µg L^{-1}]	120	700	1350
Chromium	[µg L^{-1}]	150	440	900
Copper	[µg L^{-1}]	110	250	500
15 (US-EPA) PAHs	[µg L^{-1}]	4	8	25

Table B.1: The "threshold" concentrations of the examined organic and inorganic substances for various classes of CDW (Bundesgesetzblatt, 2021).

B.2 SELECTED RANGE AND TUNED HYPERPARAMETERS OF ALGORITHMS

	LASSO	Ridge	RF			ET		
	L1	L2	n_est	max_features	max_depth	n_est	max_features	max_depth
Search Space	[0.001,10,0.005]	[0.001,10,0.005]	[1,5,251]	[1,1,18]	[1,1,10]	[1,5,251]	[1,1,18]	[1,1,10]
Sulfate	8.6	8.7	**30**	**10**	**5**	180	17	6
Vanadium	4.3	8.7	75	10	6	**180**	**17**	**5**
Copper	5.2	0.7	145	14	3	**170**	**15**	**4**
Chromium	6.4	8.6	180	17	5	**90**	**15**	**6**
15 (US-EPA) PAHs	0.6	8.6	**15**	**14**	**5**	10	17	7

L1: LASSO penalty factor; L2: Ridge penalty factor; RF: Random Forest; ET: Extremely randomized Trees; n_est: number of trees in the forest; max_features: number of features for the best split; max_depth: maximum depth of the tree.
Bold values represent the optimized hyperparameters of each specific model.

Table B.2: Tuned Hyperparameters for LS=2 and 4 predictive ML models.

B.3 OVERALL PERFORMANCE OF SEQUENCE-TIMEPOINT, HYBRID AND EARLY-STAGE INPUT MODELS TO PREDICT CONCENTRATION AT LS=10

	Sequence-timepoint				Hybrid				Early-stage			
	R^2_{train}	R^2_{CV}	R^2_{test}	$NRMSE_{test}$	R^2_{train}	R^2_{CV}	R^2_{test}	$NRMSE_{test}$	R^2_{train}	R^2_{CV}	R^2_{test}	$NRMSE_{test}$
Sulfate	0.95	0.38	0.92	0.31	0.94	0.39	0.83	0.47	0.98	0.01	0.77	0.54
Vanadium	0.87	-0.02	0.92	0.15	0.87	-0.02	0.91	0.15	0.90	0.05	0.66	0.30
Chromium	0.78	-0.59	0.72	0.11	0.74	-0.36	0.71	0.10	0.96	-2.21	0.36	0.16
Copper	0.99	-2.47	0.75	0.28	0.99	-2.47	0.74	0.29	0.99	-2.76	0.66	0.33
15(US-EPA) PAHs	0.99	-0.60	0.99	0.09	0.99	-0.60	0.99	0.10	0.99	-0.27	0.98	0.15

Table B.3: Summary of key metrics for each model.

B.4 DISTRIBUTION OF THE PH, ELECTRICAL CONDUCTIVITY, AND DOC IN THE DATA SET

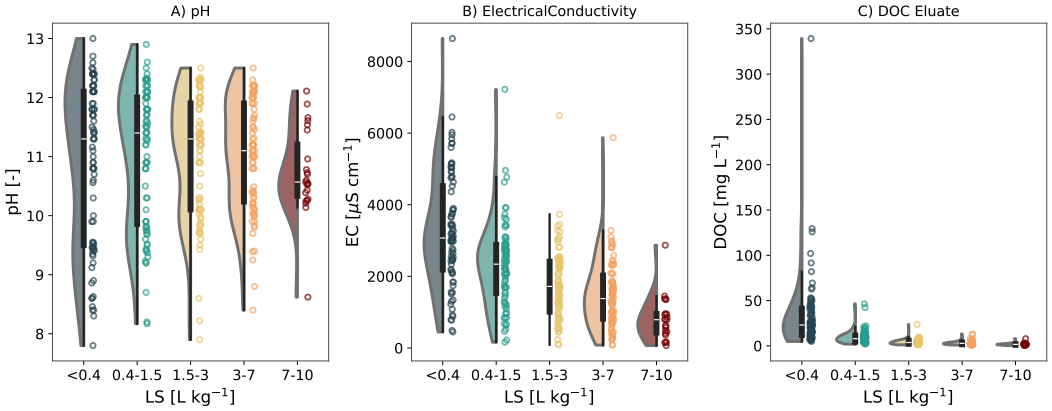


Figure B.1: Kernel probability density for distribution of pH, Electrical Conductivity, and DOC at different LS ranges. Note that the box inside the violin shapes represents the minimum, first quartile, median (white dot inside the box), third quartile, and maximum data value.

B.5 SCHEMATIC OF RANDOM FOREST ALGORITHM

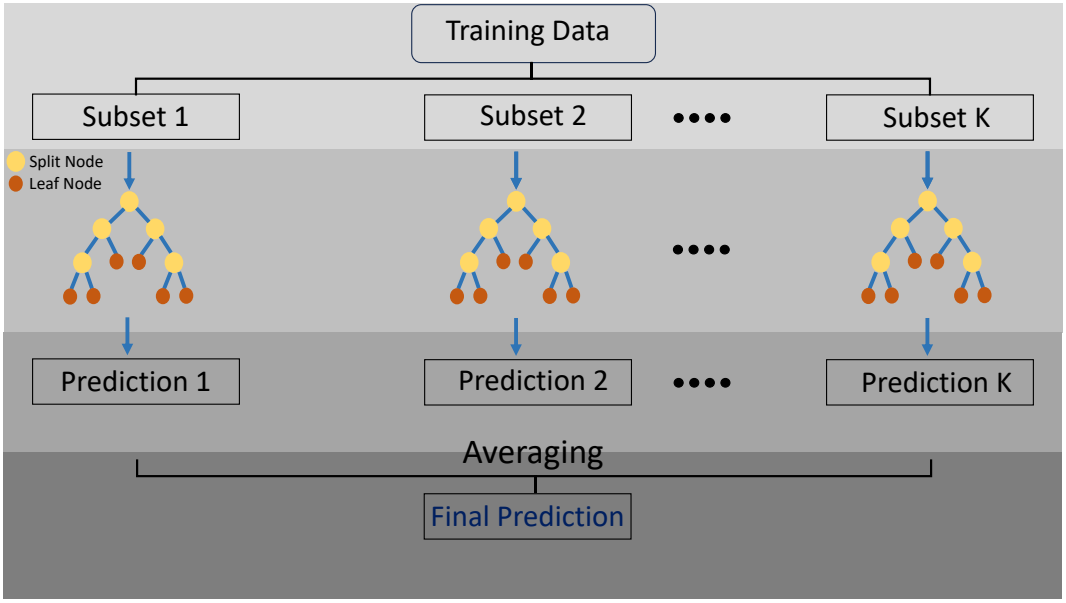


Figure B.2: Schematic of the Random Forest algorithm based on the averaging method.

B.6 MODEL VALIDATION: R^2 DISTRIBUTION

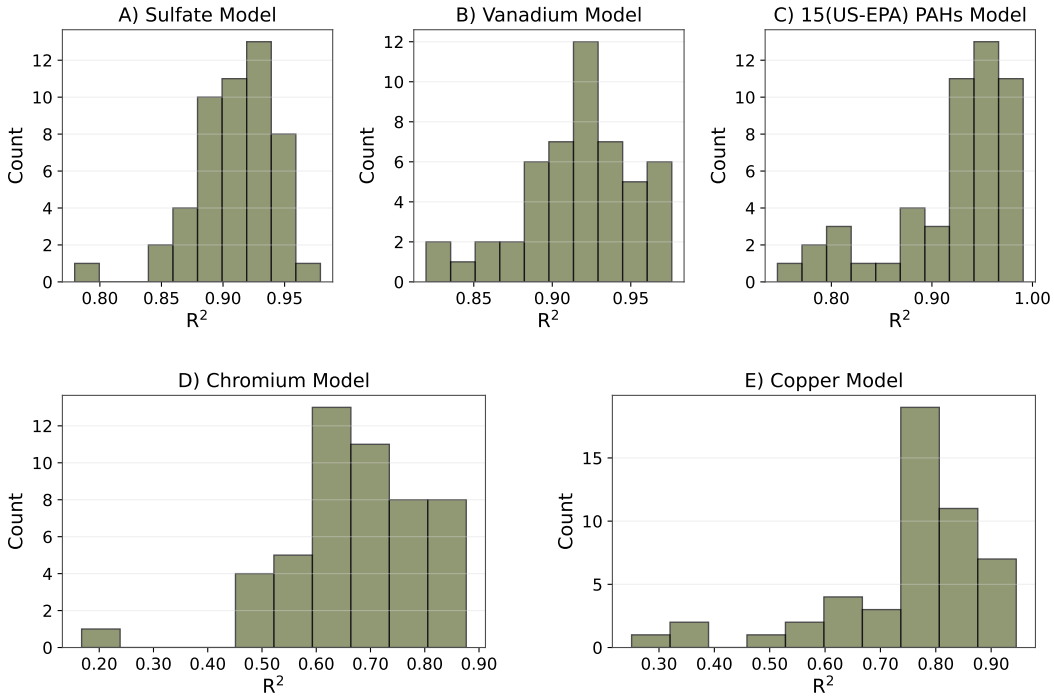


Figure B.3: Distribution of R^2 scores from repeated K-Fold Cross-Validation (5 folds, 10 repetitions) for 50 trained models at LS=2 and 4. Results are shown for A) Sulfate, B) Vanadium, C) 15 US-EPA PAHs, D) Chromium, and E) Copper.

B.7 CUMULATIVE CONCENTRATIONS PREDICTED BY ML MODELS

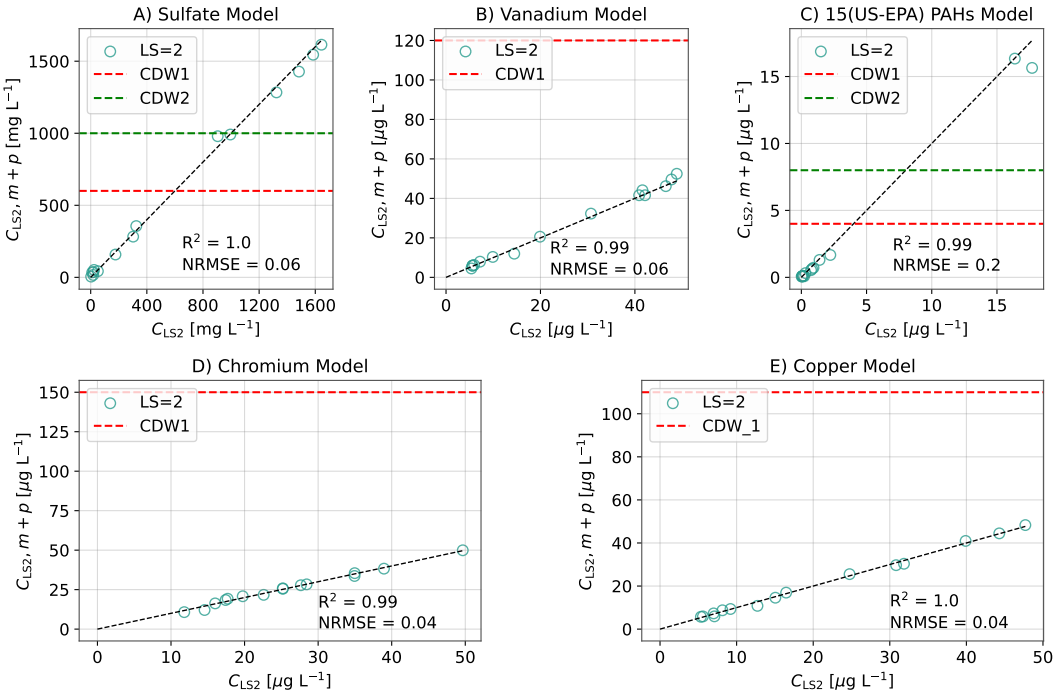


Figure B.4: Cumulative concentrations measured at LS=2 (C_{LS2}) compared to model predictions ($C_{LS2,m+p}$) using measured data at LS=0.3 and 1. The analysis is based on the best-available ML models for Sulfate, Vanadium, 15 (US-EPA) PAHs, Chromium, and Copper. Horizontal dashed lines indicate the upper bound of the respective CDW material category (see Table B.1)

B.8 PROGRAMMING ENVIRONMENT AND COMPUTER SYSTEM USED TO DEVELOP AND ANALYSE THE MODELS

The ML algorithms used for developing and optimizing surrogate models were implemented using Python 3.9.7, while the execution of the original numerical model was carried out using Julia 1.9 within the Visual Studio Code environment. Coding, learning, and analysis of the ML models took place on an Apple computer with the following specifications: Apple M1 Processor, eight-core CPU @ 3.2 GHz, and 16.0 GB RAM, running Apple's macOS operating system. The development and evaluation of ML models were facilitated using the scikit-learn package (Pedregosa et al., 2012). To ensure the reproducibility of random numbers, a Python function (random.seed) was utilized.

B.9 BOOTSTRAPPING

Bootstrapping is a metric that uses random sampling of a data set with replacement (e.g., imitating the sampling approach) and falls under the more general class of resampling methods (Efron & Tibshirani, 1994). Bootstrapping allocates accuracy measurements (bias, variance, confidence intervals, prediction error) to sample assessments. Moreover, in the Ensemble approach, each decision tree uses several subsets of data from the training sample chosen randomly with replacement to reduce the variance of each decision tree.

B.10 RANDOM-SEARCH METHOD

Random search is a widely adopted technique for optimizing the hyperparameters of a Machine Learning (ML) model. It entails randomly sampling the hyperparameter space and evaluating the model's performance at each sampled point (Bergstra & Bengio, 2012). This method is versatile and can be applied to any ML model requiring hyperparameter optimization (Feurer et al., 2015). To employ random search effectively, one must initially define the hyperparameter search space, which encompasses the range of feasible values for the hyperparameters. This search space is typically determined using a combination of domain knowledge and trial-and-error approaches. Next, the number of iterations and hyperparameter combinations to evaluate at each iteration must be specified. During each iteration, a random selection of hyperparameter values is drawn from the defined search space, and the model's performance is assessed using these values. The best set of hyperparameters is then identified by comparing the model's performance across iterations.

C

APPENDIX FOR CHAPTER 3

C.1 EMPIRICAL MODELS FOR K_d VALUES OF HEAVY METALS IN SOIL

The distribution coefficient (K_d) is a critical parameter in understanding the mobility and bioavailability of metal contaminants in soils. K_d values are influenced by various soil properties, notably pH and organic carbon (OC) content. To predict K_d values for different metal contaminants, we employed a set of empirical models. These models, derived from extensive laboratory data, relate K_d to pH and OC through logarithmic and linear transformations. The metals considered in this study include copper (Oorts, 2013), cadmium (Smolders & Mertens, 2013), manganese (Sheppard et al., 2009), nickel (Sheppard et al., 2009), and zinc (Mertens & Smolders, 2013). The empirical models for each metal are expressed as follows:

1. **Copper (Cu)**
$$\log K_\mathrm{d} = 0.4 + 0.23 \cdot \mathrm{pH} + 0.65 \cdot \log_{10}(\%\mathrm{OC}) \tag{C.1}$$

2. **Cadmium (Cd)**
$$\log K_\mathrm{d} = -1.04 + 0.55 \cdot \mathrm{pH} + 0.7 \cdot \log_{10}(\%\mathrm{OC}) \tag{C.2}$$

3. **Manganese (Mn)**
$$\log K_\mathrm{d} = -0.330 + 0.457 \cdot \mathrm{pH} \tag{C.3}$$

4. **Nickel (Ni)**
$$\log K_\mathrm{d} = 0.718 + 0.233 \cdot \mathrm{pH} + 0.0148 \cdot \%\mathrm{OC} \tag{C.4}$$

5. **Zinc (Zn)**
$$\log K_\mathrm{d} = -1.77 + 0.66 \cdot \mathrm{pH} + 0.79 \cdot \log_{10}(\%\mathrm{OC}) \tag{C.5}$$

As demonstrated in Figure C.3, the developed ensemble surrogate model shows a range of applicability based on soil pH and organic carbon content (OC) across different soil types. The heatmap plots illustrate how K_d values for various metals (Copper, Cadmium, Manganese, Nickel, and Zinc) vary with changes in pH and OC content. These models are applicable to different soil types, including mineral and organic soils, under various pH conditions. Specifically, the models are effective for predicting the leaching behavior of these metals in acidic soils and, in some instances, neutral soils.

C.2 DISTRIBUTION COEFFICIENTS K_d OF ORGANIC COMPOUNDS

The distribution coefficient (K_d) of an organic chemical can vary widely across different soils or sediments due to variations in the properties of the sorbent. However, it is well-established that for many organic chemicals, especially neutral hydrophobic ones, the distribution coefficient is directly proportional to the amount of organic matter present

(Grathwohl, 1990; Doucette, 2003). This relationship is captured by the normalized organic carbon partition coefficient (K_{OC}), defined by the equation:

$$K_{OC} = \frac{K_d}{OC}, \tag{C.6}$$

where K_d is the distribution coefficient, and OC is the organic carbon content of the sorbent, expressed as grams of organic carbon per gram of soil [g OC g^{-1} soil].

The applicability of the developed ensemble surrogate model for organic compounds is determined by the predefined K_d range (i.e., 0.1-25 [L kg^{-1}]) in different range of K_{OC} and organic carbon content, as illustrated in Figure C.4. This range demonstrates that the model can be suitable to simulate the leaching behavior of organic compounds with K_{OC} values within this range across various soil types.

For instance, compounds such as acetophenone, with a K_{OC} of approximately 42 [L kg^{-1}] (Khan et al., 1979), 1,2-dibromoethane with a K_{OC} of 44 [L kg^{-1}] (Kenaga & Goring, 1980), and benzene, with a K_{OC} of around 80 [L kg^{-1}] (Hodson & Williams, 1988), fall well within the applicable range. This suggests that the ensemble surrogate model in the predefined K_d range can be applied to predict the organic compound leaching behavior with low K_{OC} values in different soil types, including low to moderate organic content. Similarly, compounds like atrazine with higher organic carbon partition coefficient (i.e., K_{OC} values around 150 [L kg^{-1}]; Kenaga, Goring, 1980), and toluene, with a K_{OC} of approximately 160 [L kg^{-1}] (Hodson & Williams, 1988), also fall within this range, indicating potential applicability in soils with varying organic content. Nitrobenzene, with a K_{OC} of about 200-400 [L kg^{-1}] (Løkke, 1984), and naphthalene with a K_{OC} of approximately 1000 [L kg^{-1}] (Løkke, 1984) are additional examples that fit within the surrogate model's defined K_d range but are limited to soils with low organic content.

C.3 PAIRWISE PROJECTIONS OF THE VIRTUAL REALITY DATASETS (VRD)

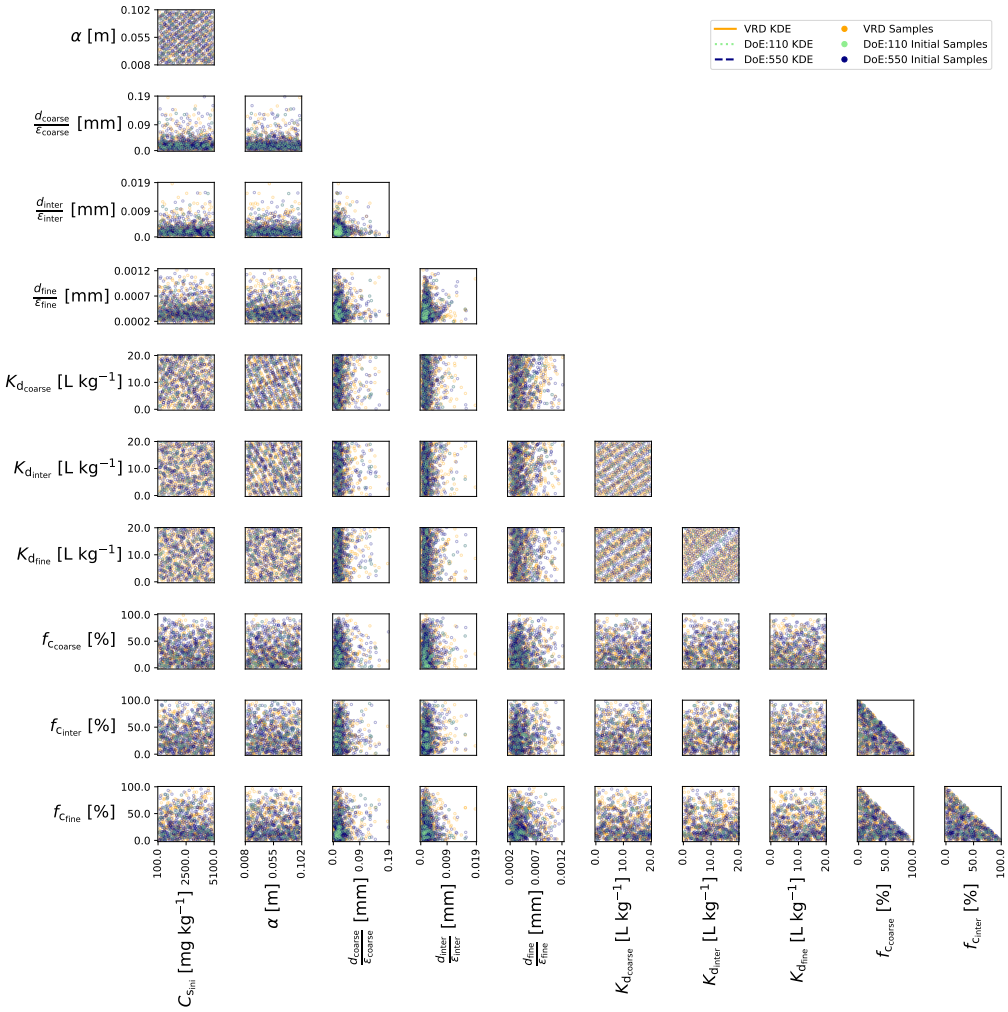


Figure C.1: Pairwise projections of surrogate model parameters used to generate Virtual Reality Datasets (VRDs) and Designs of Experiments (DoE) based on Halton sequence sampling. The VRDs are depicted in orange, while DoE datasets are represented in green for 110 initial samples and navy for 550 initial samples. These scatter plots illustrate the coverage of the parameter space for each pair of parameters, demonstrating the spread and interaction within the predefined parameter space. Notably, the ratios of d to intra-particle porosity, being divisions of two uniform distributions, cover certain regions of the parameter space uniquely. Additionally, the last three parameters represent volume fractions whose total must sum to 100%, resulting in simplex shape.

C.4 CORRELATION COEFFICIENT ANALYSIS FOR POSTERIOR PARAMETER DISTRIBUTION

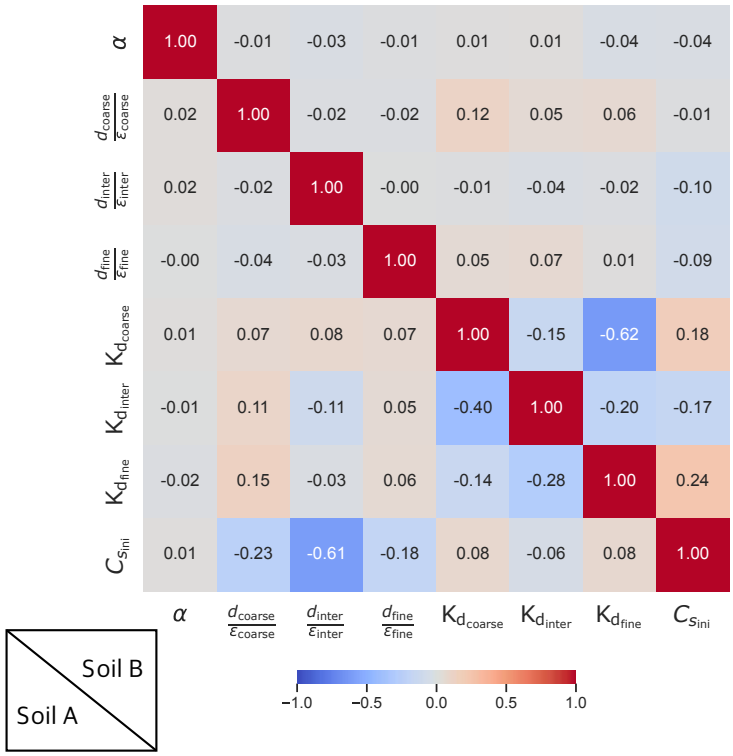


Figure C.2: Correlation coefficients between posterior parameters. The upper and lower triangles represent correlation coefficients for Soil A and Soil B, respectively.

C.5 RANGE OF APPLICABILITY OF DEVELOPED ENSEMBLE SURROGATE MODEL FOR SELECTED HEAVY METALS

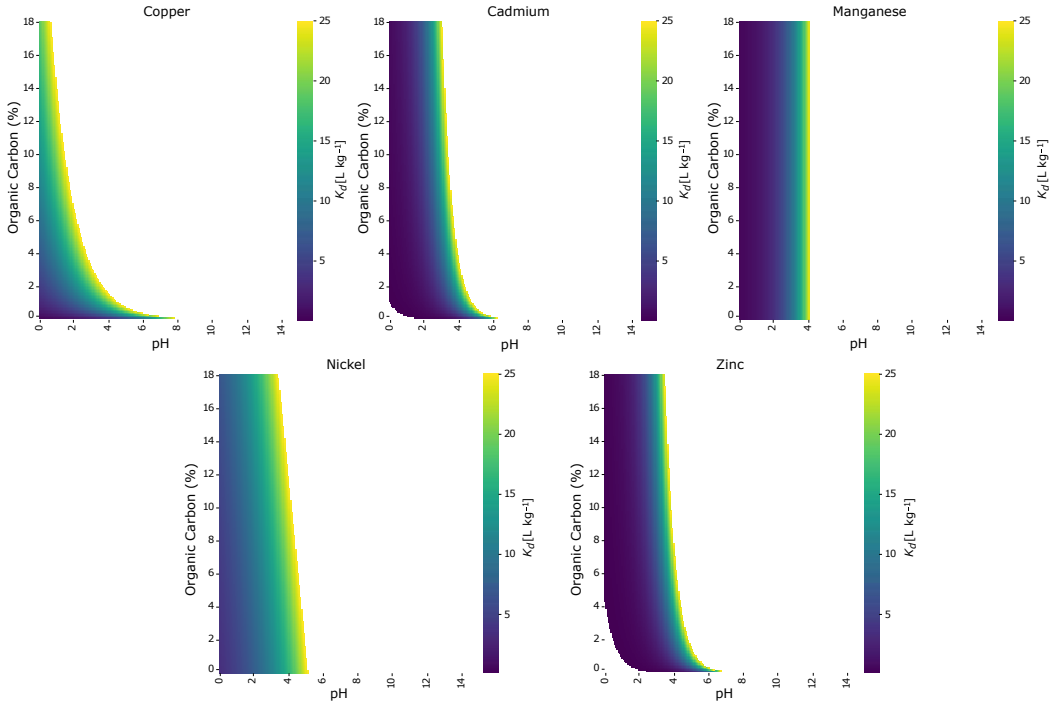


Figure C.3: Heatmap plots illustrating the range of applicability of the developed ensemble surrogate model based on pH and organic carbon (%OC). The plots depict the relationship between soil pH, organic carbon content, and distribution coefficients (K_d) for various metals (Copper, Cadmium, Manganese, Nickel, and Zinc). Each subplot demonstrates how K_d values change with variations in pH and organic carbon percentage for a specific metal. These relationships are derived from empirical equations tailored to each metal (refer to C.1 for more details). The color intensity in each heatmap indicates the magnitude of the distribution coefficient, providing a visual representation of the interactive effects of soil pH and organic carbon on metal retention.

C.6 RANGE OF APPLICABILITY OF DEVELOPED ENSEMBLE SURROGATE MODEL FOR ORGANIC COMPOUNDS

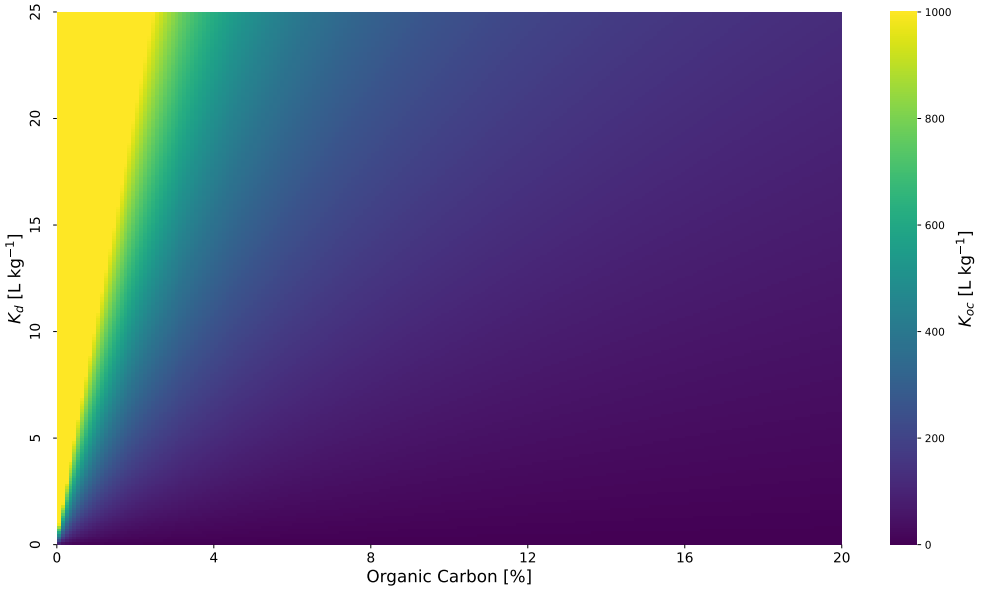


Figure C.4: Heatmap depicting the relationship between organic carbon content and the distribution coefficient (K_d) for various values of the organic carbon partition coefficient (K_{OC}). This heatmap illustrates how K_{OC} values change with variations in organic carbon percentage and K_d. The color intensity in the heatmap represents the magnitude of K_{OC}, providing a visual representation of the interactive effects between organic carbon content and K_d. This visualization helps in understanding the applicability of the ensemble surrogate model for predicting the behavior of organic compounds in diverse soil environments.

C.7 CONSTANT PARAMETERS FOR ADE-IPD MODEL (DIN 19528)

Parameter [Unit]	Value	Description
Saturation time [h]	$t_{sa} = 2$	Time required for the system to saturate the column with fluid.
Equilibrium Time after Saturation [h]	$t_{eq} = 0$	Time required for the system to reach equilibrium after saturation with fluid.
Percolation time [h]	$t_{pe} = 5$	Time that solid phase is in contact with fluid during percolation.
Column Length [m]	$x_{col} = 0.3$	Length of the column in the system.
Dry Solid Density [kg L^{-1}]	$\rho_s = 2.60$	Density of the solid material in the system.

Table C.1: Constant parameters used to evaluate ADE-IPD model based on German column leaching test standard (DIN 19528, 2009)

C.8 IMPACT OF SURROGATE MODEL COMBINATION ORDER ON STACKING MODEL PERFORMANCE

	DoE 110	DoE 550
Max Mahalanobis	118	200
Min Mahalanobis	140	200

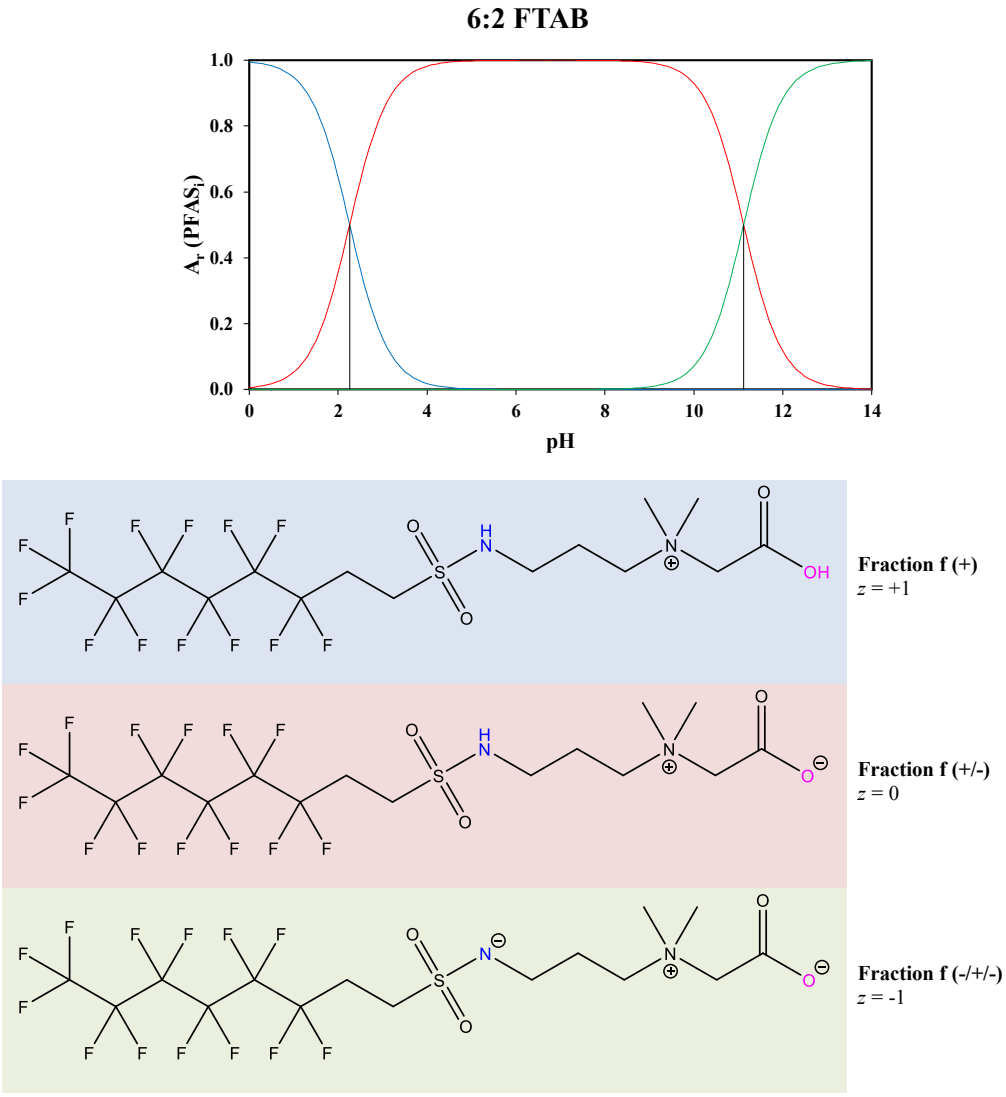
Note: The Stacking Model for the DoE with 550 initial samples did not reach the stop criteria, hence resulting in the utilization of 200 surrogate models.

Table C.2: Number of individual surrogate models utilized in stacking methods to reach the stop criteria (RRMSE= 10%), based on maximum and minimum Mahalanobis distances between each surrogate model and the previously stacked model.

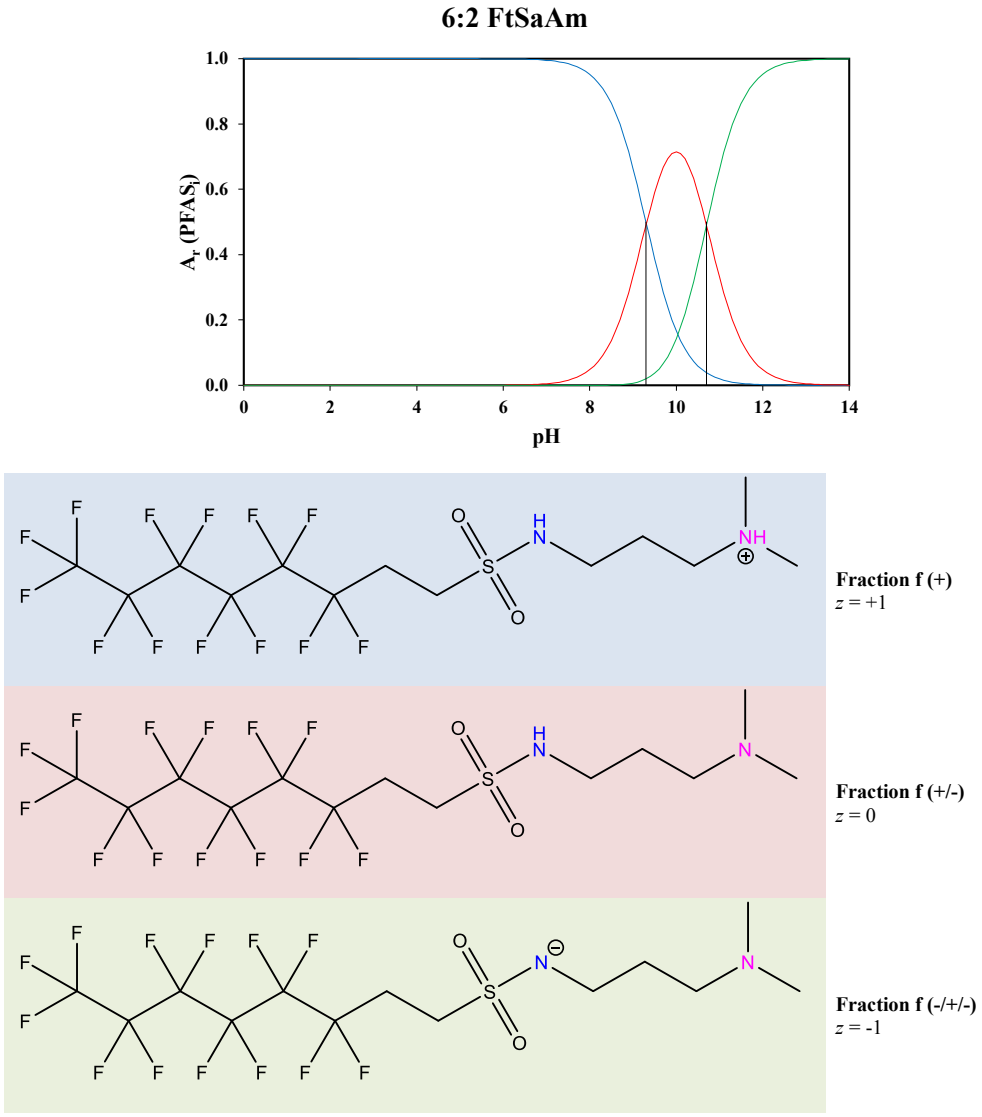
D

APPENDIX FOR CHAPTER 4

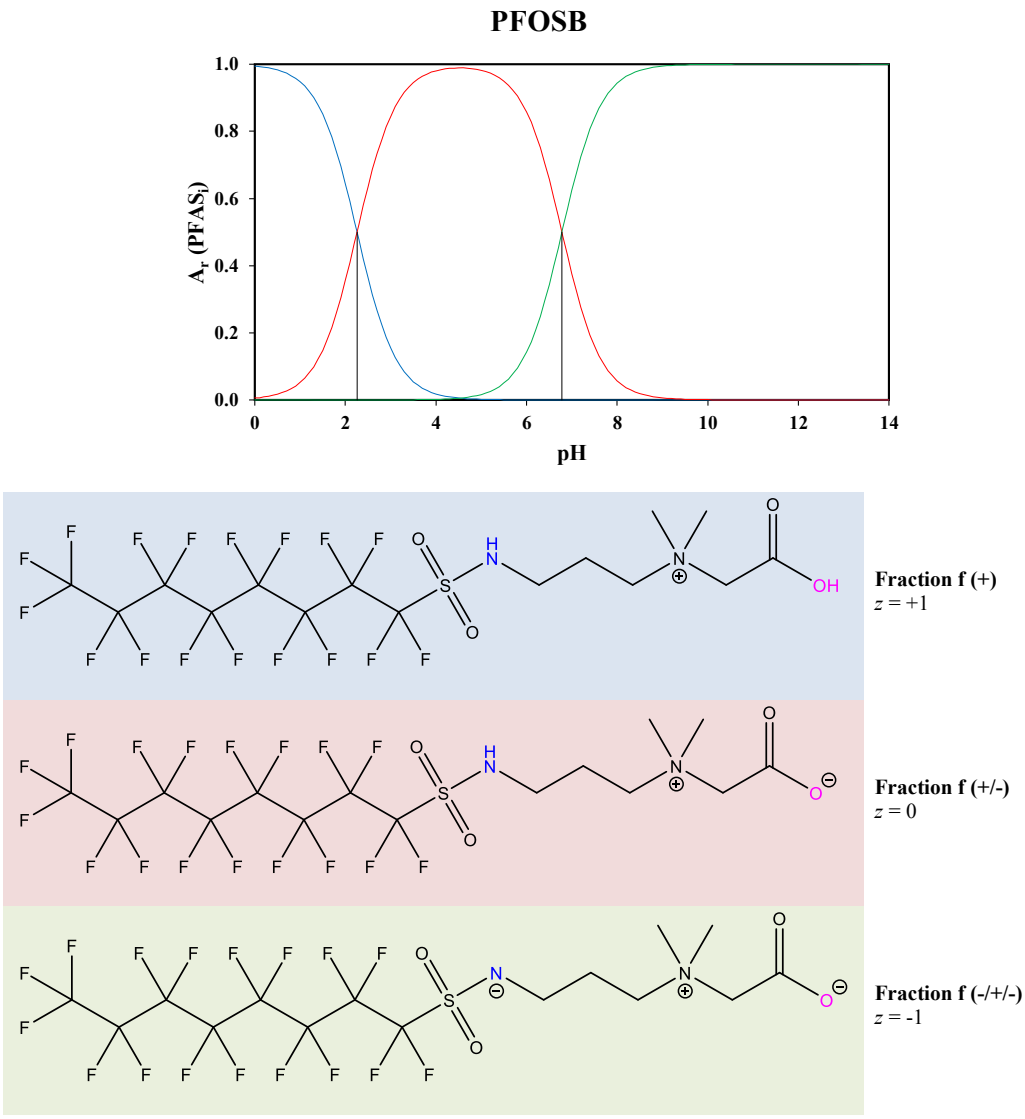
D.1 : PH-SPECIATION DIAGRAM FOR SOME REPRESENTATIVE PFAS SPECIES



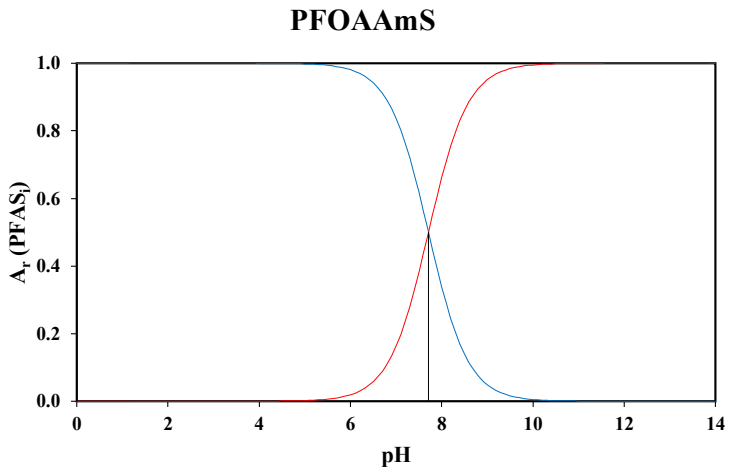
Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*

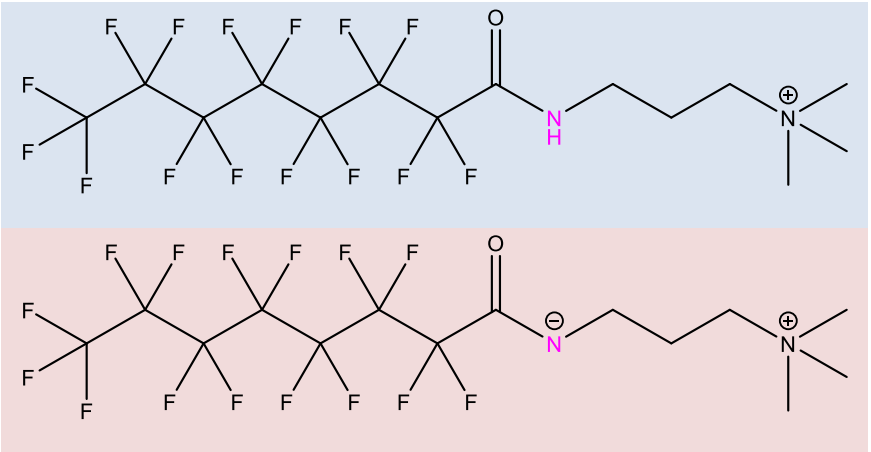


Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*

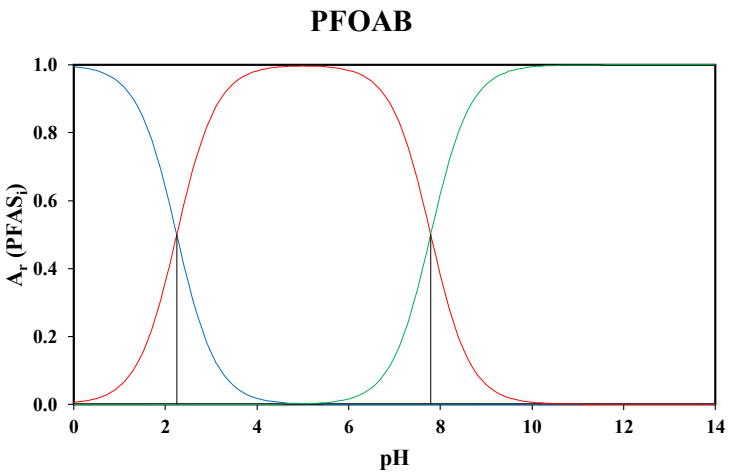


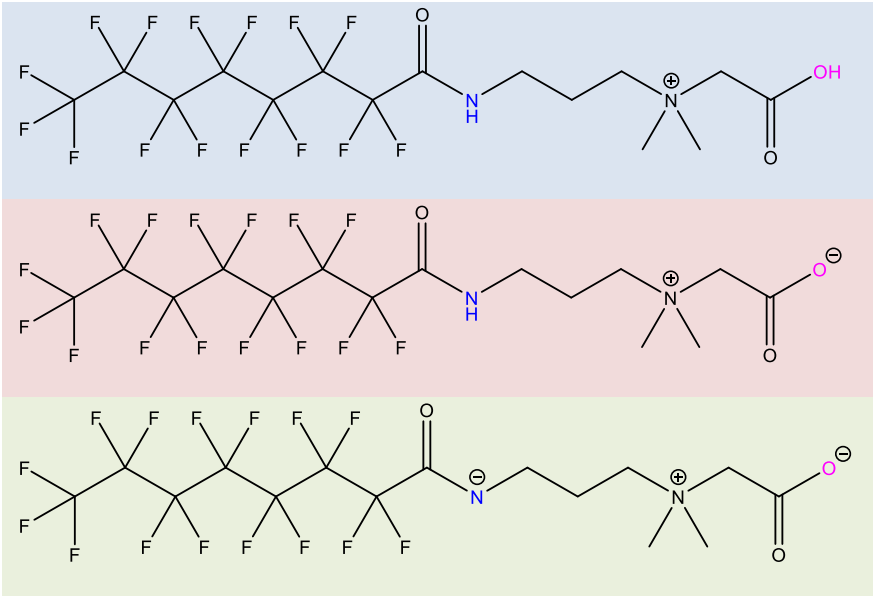
Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*



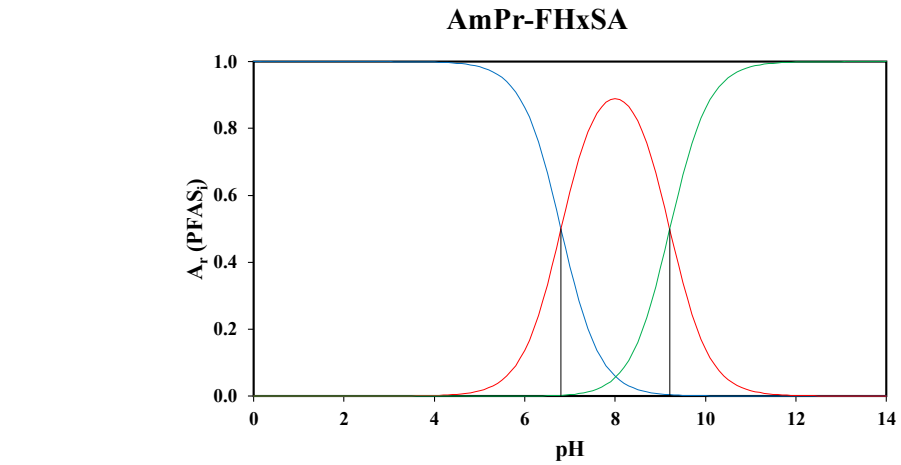


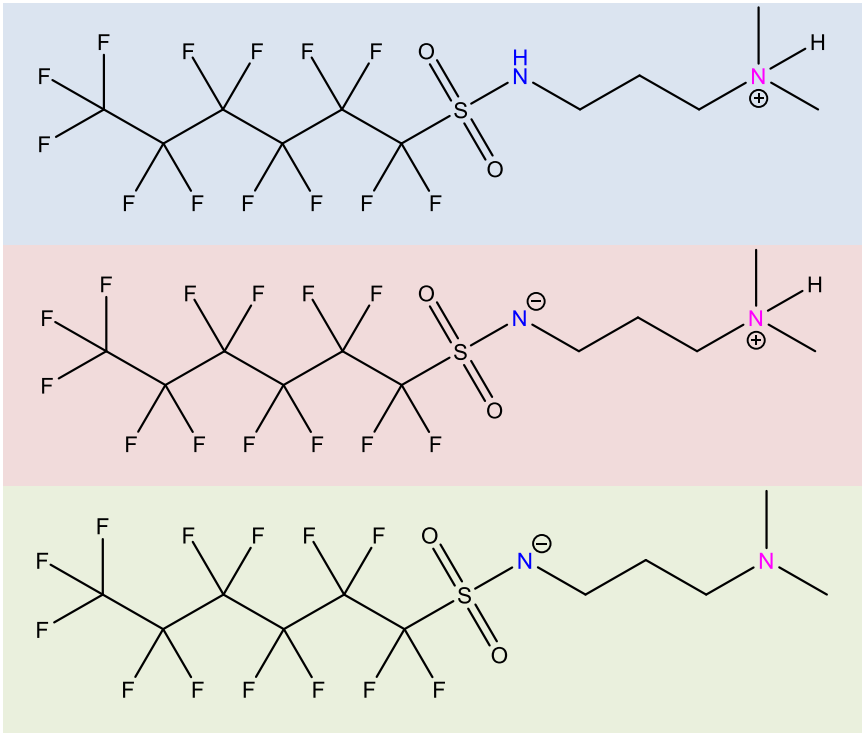
Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*



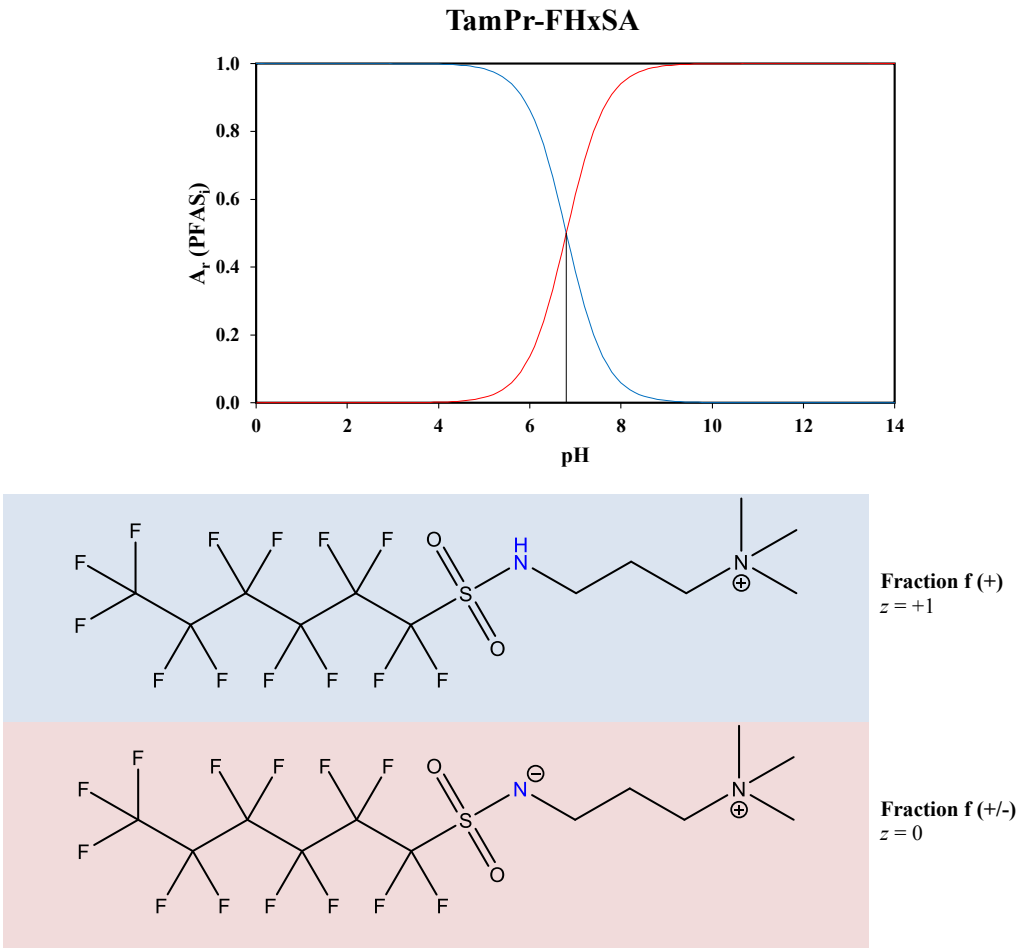


Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*



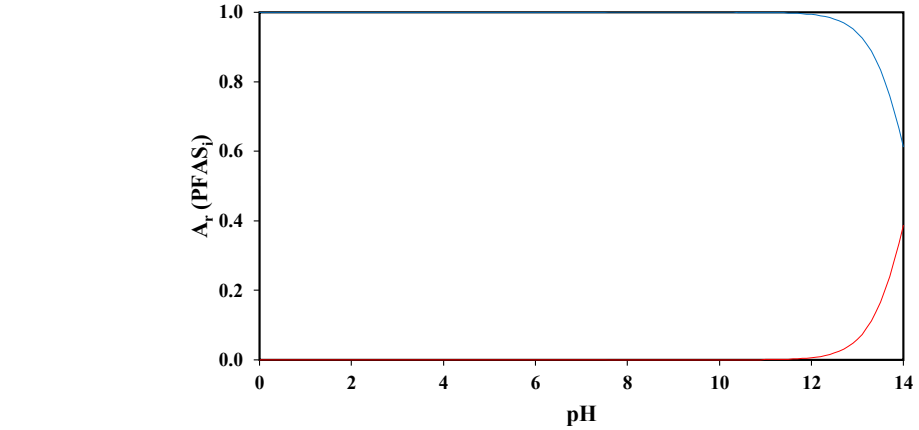


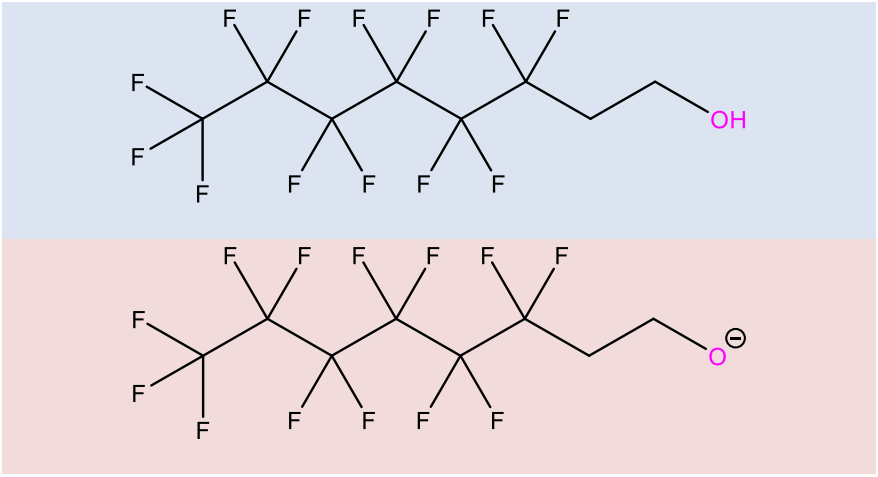
Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*



Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*

6:2 FTOH

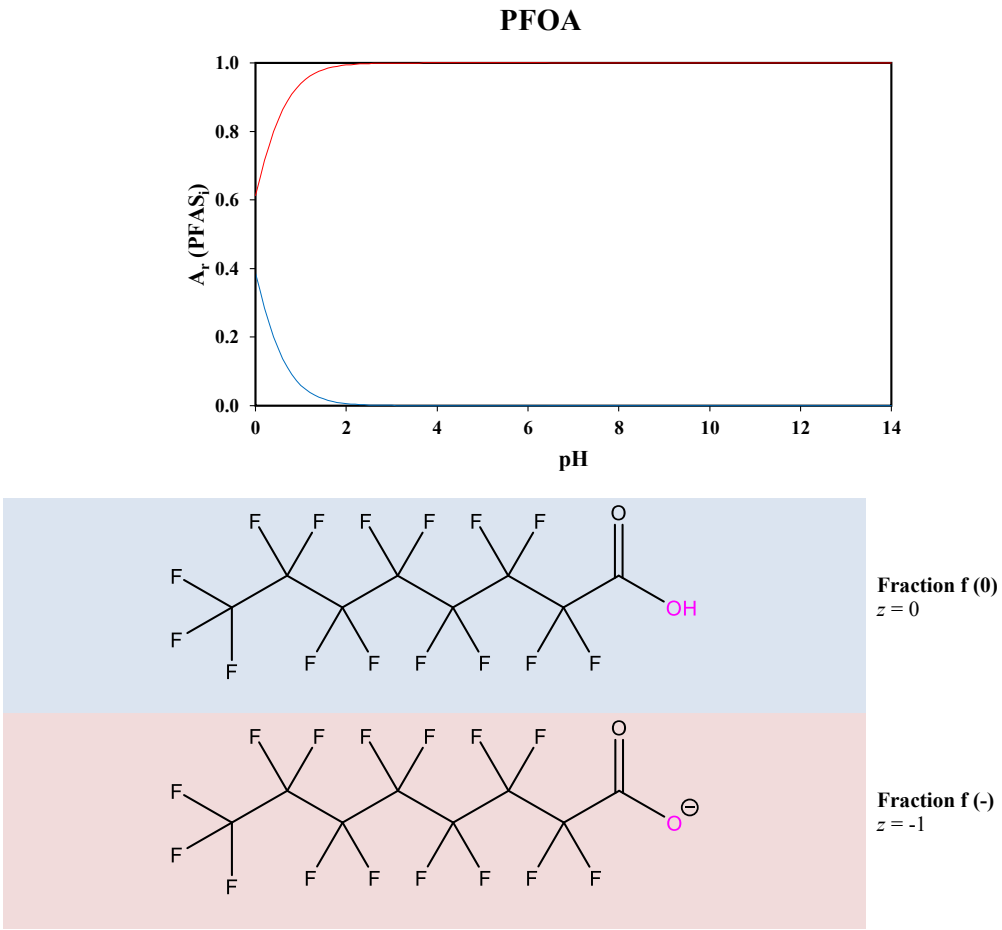




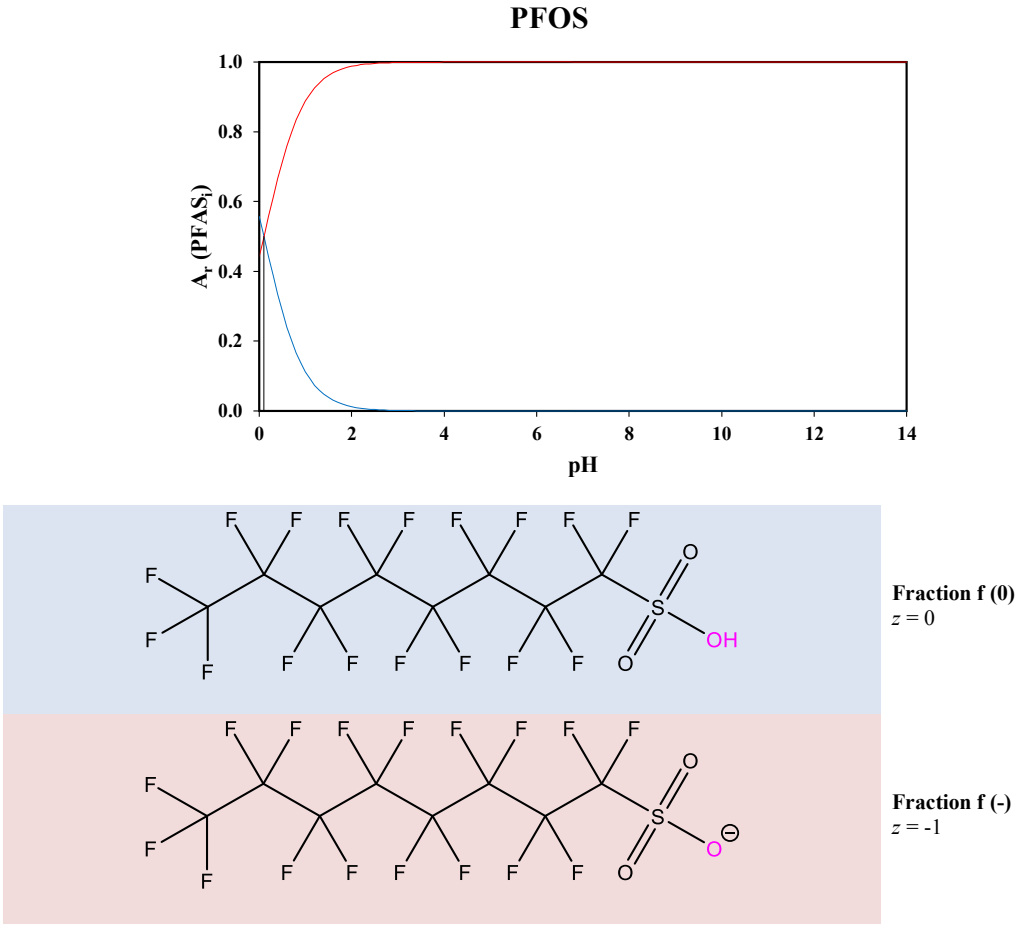
Fraction f (0)
$z = 0$

Fraction f (-)
$z = -1$

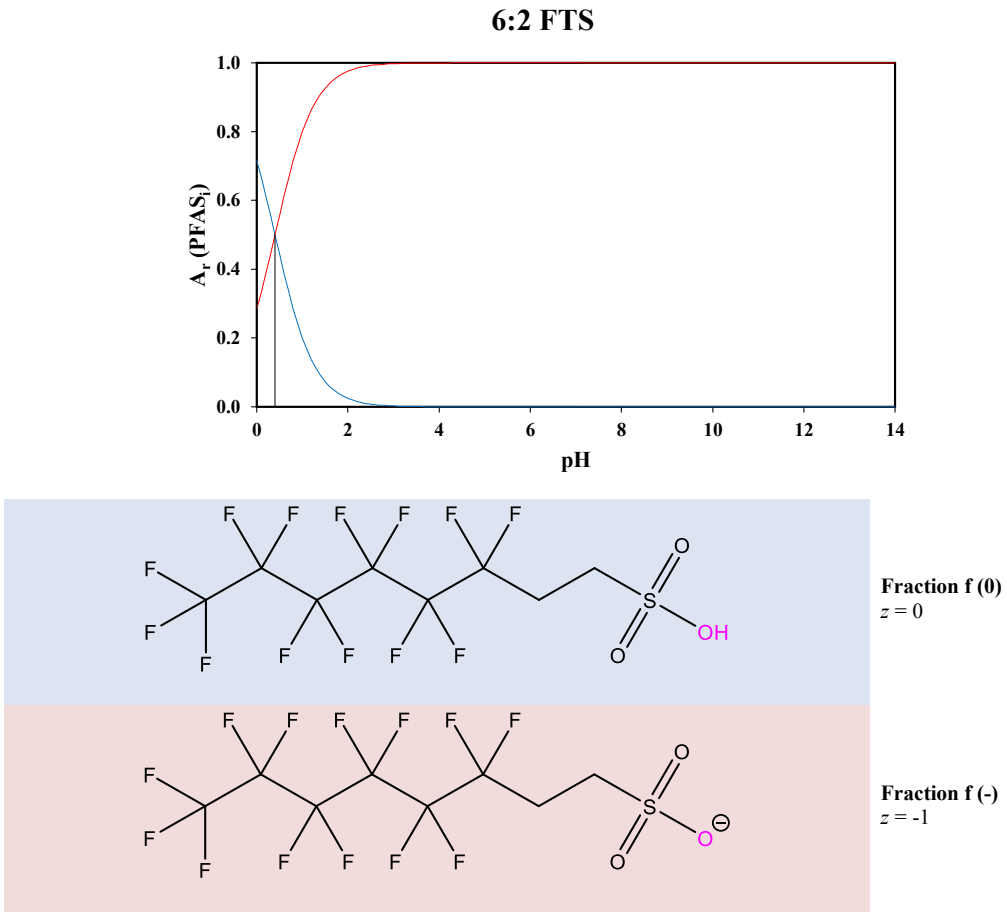
Z indicates molar net charge [-]; A_r *($PFAS_i$)* indicate the relative abundance of each PFAS specie *i*



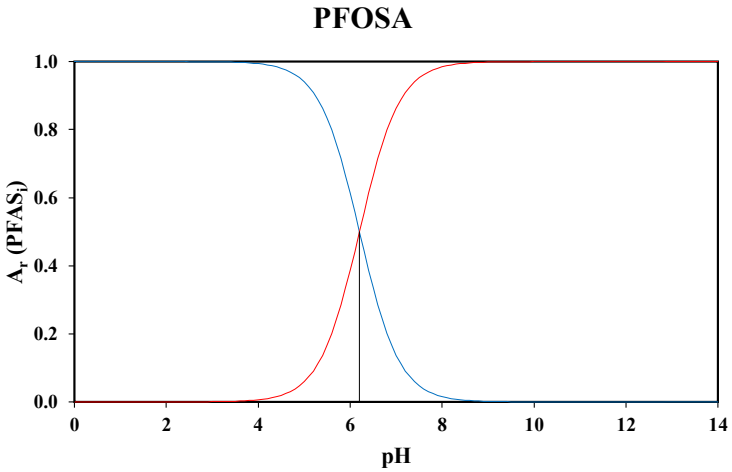
Z indicates molar net charge [-]; A_f *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*

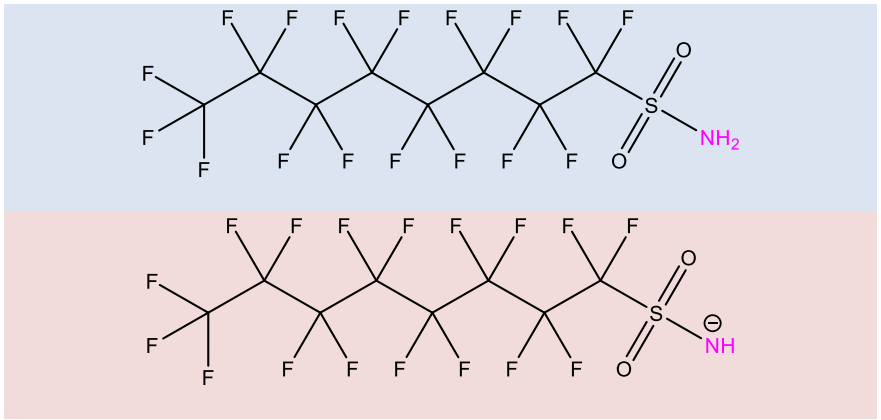


Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*



Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*

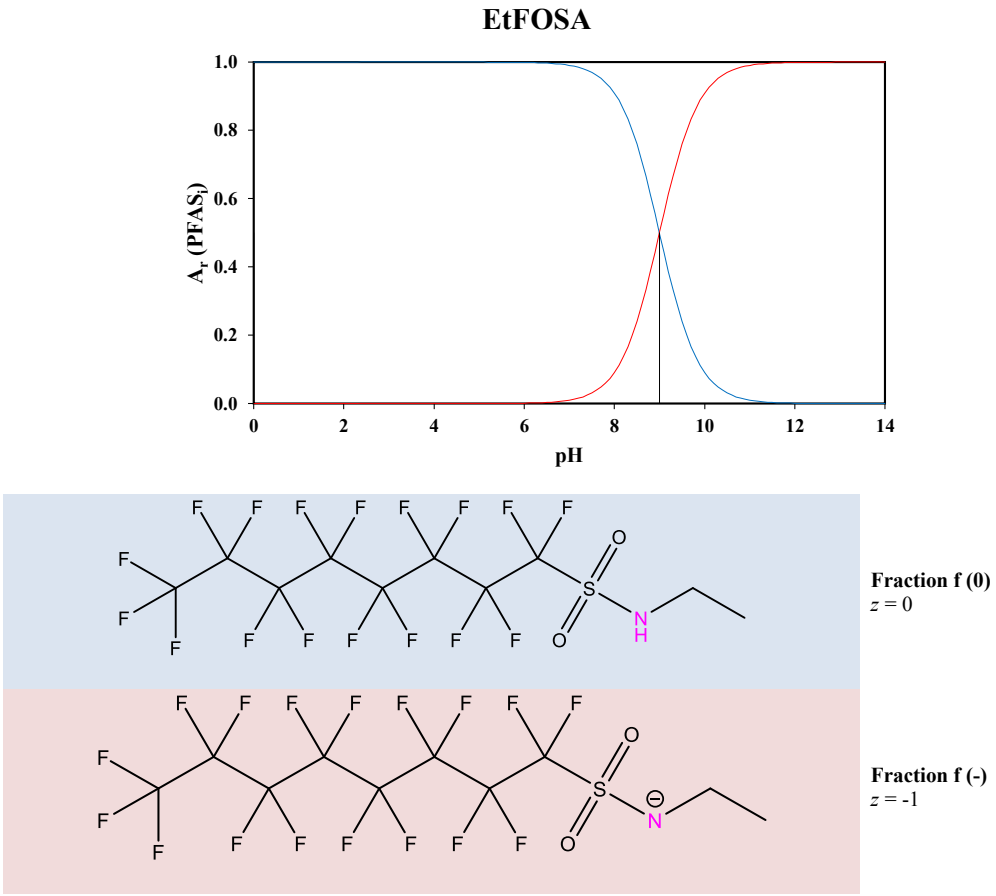




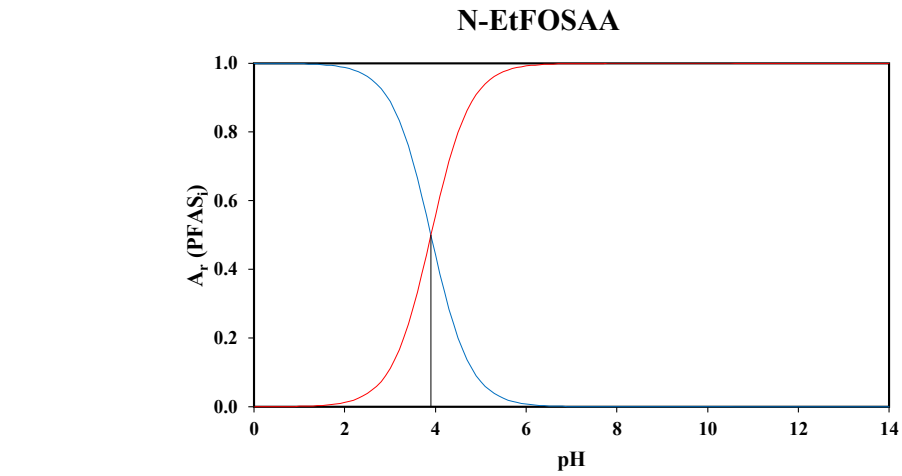
Fraction f (0)
$z = 0$

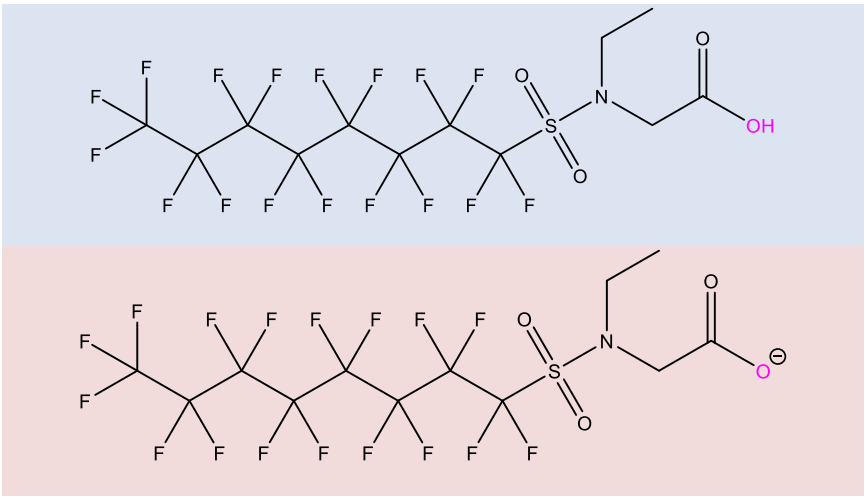
Fraction f (-)
$z = -1$

Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*



Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*

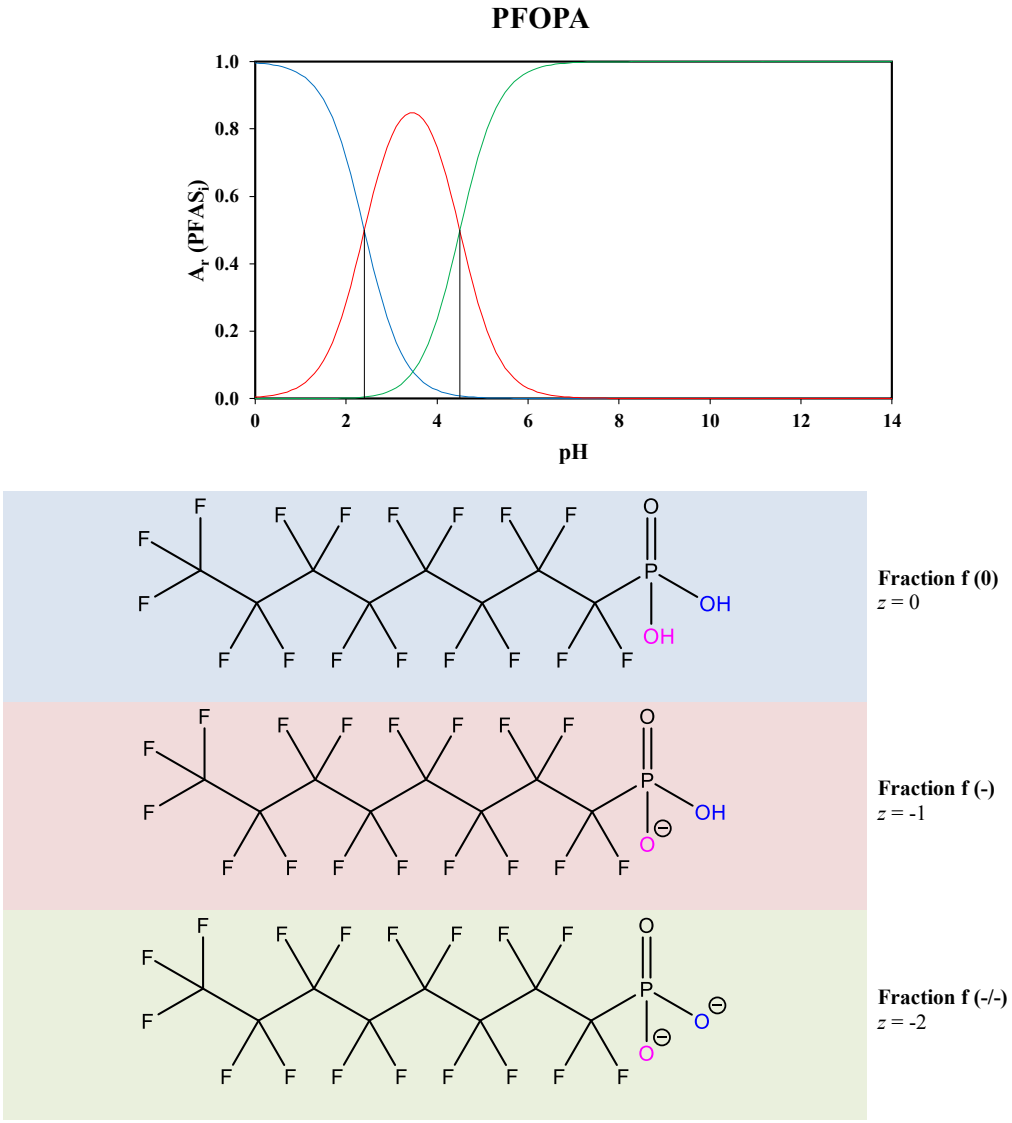




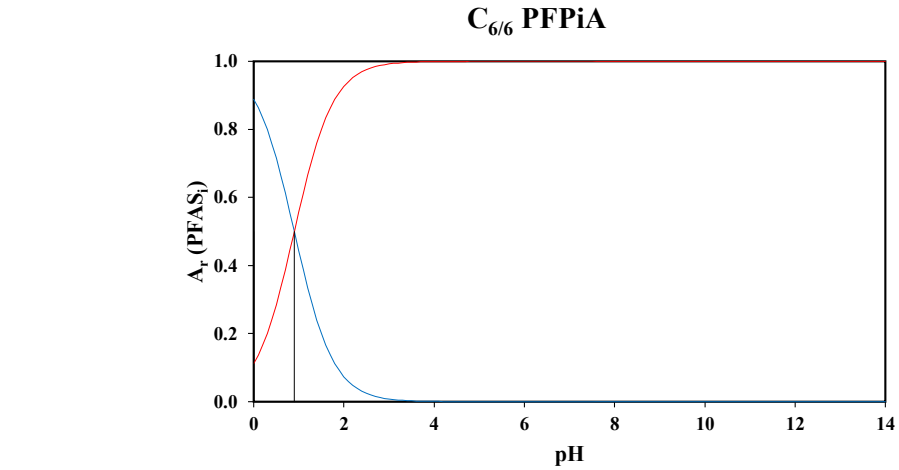
Fraction f (0)
$z = 0$

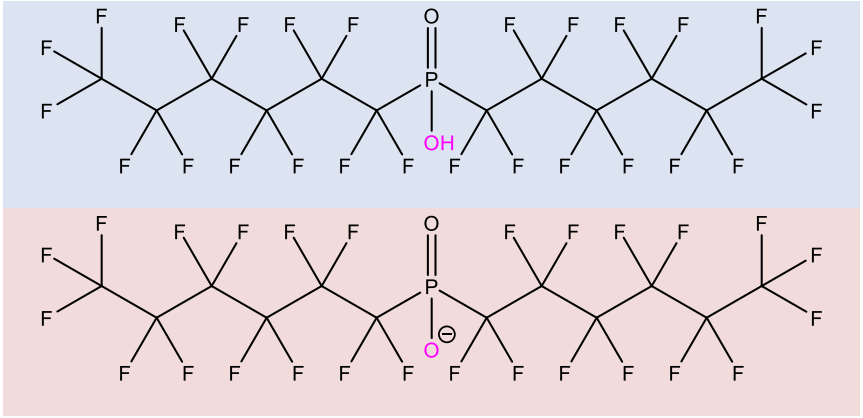
Fraction f (-)
$z = -1$

Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*



Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*





Z indicates molar net charge [-]; A_r *(PFAS$_i$)* indicate the relative abundance of each PFAS specie *i*

D.2 : ADDITIONAL INFORMATION ON THE DERIVATION OF K_d DATA FROM LITERATURE STUDIES

The solid-liquid distribution coefficient K_d [L kg^{-1}] is defined as the ratio between the sorbed pollutant concentration in the solid phase C_S [mg kg^{-1}] to the non-sorbed pollutant concentration in the water phase C_W [mg L^{-1}] under equilibrium conditions, as shown in Equation D.1:

$$K_d = \frac{C_S}{C_W}. \tag{D.1}$$

K_d is therefore a parameter suitable to assess the concurrent equilibria for pollutants between solid and liquid phases in the environment, and a relevant input parameter used in contaminant transport modeling (Guo et al., 2020), with higher K_d values attributed to higher sorption to soil particles and, therefore, a higher retardation on contaminant transport to groundwater.

Specifically, for PFAS studies assessing sorption in soils, K_d values have a wide range of variability. For instance, K_d values ranging from < 0.1 to 10 L kg^{-1} have been reported for short-chained PFAS such as TFA and PFBA (Nguyen et al., 2020; Richey et al., 1997), while K_d values ranging from 90 to 5, 700 L kg^{-1} have been reported for long-chained PFAS such as PFDoA (Nguyen et al., 2020; Campos Pereira et al., 2018). K_d values reported for PFOA range from < 1 to 130 L kg^{-1}(Nguyen et al., 2020; Fabregat-Palau et al., 2021). These discrepancies in observed K_d values are a result of the different soil properties, especially the differing organic carbon content (C_{org}). As such, K_d values are usually normalized to the soil fraction of organic carbon (f_{OC}), thus defining the organic-carbon normalized sorption coefficient K_{OC} [L kg OC^{-1}] as shown in Equation D.2:

$$K_{\text{OC}} = \frac{K_d}{f_{\text{OC}}}. \tag{D.2}$$

The normalization of K_d to K_{OC} is suitable to reduce K_d variability for compounds whose main sorption domain is the soil organic carbon content (C_{org}) (Fabregat-Palau et al., 2021). In the case of PFAS, however, low but significant sorption is observed on pure inorganic minerals such as kaolinite and soils with low C_{org} (Xiao et al., 2011; Knight et al., 2019), which may lead to a plausible overestimation of K_{OC} in low soil C_{org} scenarios. Although these overestimations may be overcome by the inclusion of a term in Equation D.2 representative of the sorption in reactive mineral components (Fabregat-Palau et al., 2021), in this work, we derived individual PFAS/sorbent K_{OC} values for each entry in the dataset of Supplementary File 1 (download it from https://hydrogeochem.geo.uni-tuebingen.de/pfas) according to Equation D.2 to assess the role of PFAS chain length and functional group on sorption.

Sorption isotherms describe the relationship between C_S and C_W over a certain concentration range. The resulting isotherm shape may be linear (where the slope, i.e., K_d, is independent of the tested concentration) or non-linear (where the slope, i.e., K_d, is dependent on the tested concentration). Non-linear isotherms are often described by the Freundlich equation (Equation D.3):

$$C_S = K_{\text{Fr}} C_W^{1/n}, \tag{D.3}$$

where K_{Fr} is the Freundlich constant [(mg kg^{-1}) / (mg L^{-1})$^{1/n}$], representative of the sorbent-sorbate sorption affinity, and $1/n$ represents the non-linearity term. For sorbents where the sorption site availability is limited with respect to the number of pollutant molecules in the system, a non-linearity term with $1/n < 1$ is often observed, and K_d decreases with increasing pollutant concentration. On the other hand, cooperative sorption (i.e., additional sorption sites provided by sorbed pollutants due to pollutant–pollutant interactions) may result in non-linear isotherms with $1/n > 1$, where K_d increases with increasing pollutant concentration. Regarding PFAS, both linear (i.e., $1/n = 1$) and non-linear isotherms (i.e., $1/n < 1$ and $1/n > 1$) have been observed (Mejia-Avendaño et al., 2020; Milinovic et al., 2015; Li et al., 2019).

Non-linear sorption isotherms may also be described by the Langmuir model (Equation D.4) when a complete saturation of sorption sites occurs. The model allows derivation of the Langmuir constant K_L [L kg^{-1}], representative of the sorbent-sorbate sorption affinity, and the maximum loading capacity Q_{MAX} [mg kg^{-1}], representative of the total number of sorption sites borne by the sorbent:

$$C_S = \frac{K_L Q_{\mathrm{MAX}} C_W}{1 + K_L C_W}. \tag{D.4}$$

Regarding PFAS, only a few studies observed sorption isotherms well described by the Langmuir model (Wei et al., 2019).

To extend the pool of K_d (PFAS) data in our dataset, we carefully examined literature studies reporting fitted isotherm data but not providing K_d. Similar approaches have been applied elsewhere (Fabregat-Palau et al., 2022, 2023, 2024; Saeidi et al., 2024; Kleineidam et al., 2002). In some of these works (Fabregat-Palau et al., 2022, 2023, 2024), K_d values were derived from both Freundlich and Langmuir models at a certain arbitrary C_W that fell in the low concentration range of the sorption isotherm, aiming to represent low environmental concentrations (Brusseau et al., 2020) and assuming that K_d data fall in the linear range of the sorption isotherm. On the other hand, recent studies have derived K_d data from literature studies resulting from fitted isotherm parameters at the same C_W for all PFAS, allowing a FAIR comparison (Saeidi et al., 2024). Nonetheless, this approach results in the derivation of K_d values for PFAS at very different activities in water, given the differences in solubility values for short- and long-chained PFAS. To address this, some other studies have derived K_d data from literature resulting from fitted isotherm parameters at the same PFAS activity in water (i.e., C_W = 10% of PFAS solubility), which allowed a better comparison of K_{d} values among sorbates (Kleineidam et al., 2002). In this work, we applied this latter approach, deriving K_d values from reported isotherm data fitted to both the Freundlich and Langmuir models at different C_W but at the same PFAS activity in water (i.e., C_W = 10% of PFAS solubility). Care was taken to consistently derive solubility values from the same source (i.e., EPISuite) (Card et al., 2017). The assessment of whether these values differed among experimental (Inoue et al., 2012) and other modeled (e.g., Cosmotherm) (Wang et al., 2011) approaches was out of the scope of this work.

D.3 : NUMBER OF SOILS AND ENTRIES USED TO DERIVE LOG K_{OC} VALUES

PFAS	Subfamily	Number of Entries	Number of References	Average log K_{OC}
TFA	PFCA	20	1	1.18 (0.37)
PFBA	PFCA	24	7	1.33 (0.61)
PFPeA	PFCA	19	6	1.48 (0.49)
PFHxA	PFCA	26	7	1.47 (0.47)
PFHpA	PFCA	24	9	1.91 (0.46)
PFOA	PFCA	187	26	2.57 (0.70)
PFNA	PFCA	34	12	2.55 (0.37)
PFDA	PFCA	30	9	3.14 (0.52)
PFUnA	PFCA	20	6	3.71 (0.48)
PFDoA	PFCA	21	5	4.50 (0.67)
PFTrA	PFCA	3	2	4.77 (0.57)
PFTeA	PFCA	2	1	5.25 (0.70)
GenX	PFECA	10	1	1.71 (0.53)
ADONA	PFECA	10	1	1.69 (0.46)
PFBS	PFSA	34	11	1.43 (0.47)
PFPeS	PFSA	12	2	1.71 (0.45)
PFHxS	PFSA	72	11	1.82 (0.46)
PFHpS	PFSA	12	3	2.62 (0.31)
PFOS	PFSA	238	32	3.07 (0.39)
PFNS	PFSA	12	2	3.39 (0.42)
PFDS	PFSA	16	3	3.86 (0.49)
PFEtCHxS	PFSA	10	1	2.47 (0.44)
8:2 Cl-PFAES	PFAES	10	1	4.05 (0.51)
FBSA	FOSA	10	1	1.71 (0.49)
FHXSA	FOSA	10	1	2.32 (0.62)
PFOSA	FOSA	16	3	3.37 (0.61)
EtFOSA	FOSA	3	1	3.39 (0.48)
N-MeFOSAA	FOSAA	5	1	3.10 (0.31)
N-EtFOSAA	FOSAA	5	1	3.33 (0.35)
4:2 FTOH	FTOH	1	1	0.93 (N.A.)
6:2 FTOH	FTOH	3	1	2.42 (0.10)
8:2 FTOH	FTOH	5	1	3.77 (0.04)
10:2 FTOH	FTOH	3	1	4.60 (0.28)
4:2 FTS	FTS	10	1	1.69 (0.49)
6:2 FTS	FTS	27	6	2.14 (0.67)
8:2 FTS	FTS	26	5	3.01 (0.71)
PFHxPA	PFPA	7	1	1.69 (0.55)
PFOPA	PFPA	7	1	2.64 (0.39)
PFDPA	PFPA	7	1	2.77 (0.48)
$C_{6/6}$ PFPiA	PFPiA	7	1	3.19 (0.49) *
$C_{6/8}$ PFPiA	PFPiA	7	1	3.56 (0.73) *
$C_{8/8}$ PFPiA	PFPiA	7	1	2.53 (0.80) *
6:2 FtSaB	Zwitterionic	22	4	2.75 (1.12)
8:2 FtSaB	Zwitterionic	5	1	4.14 (1.06)
10:2 FtSaB	Zwitterionic	2	1	4.54 (0.34)
6:2 FtSaAm	Cationic	1	1	5.93 (N.A.) *
PFOSB	Zwitterionic	5	1	2.46 (0.76)
PFOAAmS	Cationic	11	3	3.31 (0.50)
PFOAB	Zwitterionic	11	3	2.30 (0.60)
AmPr-FHSA	Cationic	10	1	3.26 (0.88)
TAmPr-FHSA	Cationic	10	1	3.47 (1.10)

Table D.1: Summary of the total entries and references for each PFAS compound, with averaged log K_{OC} values (Equation D.2). Brackets indicate standard deviation; * marks outliers excluded from training. N.A.: Not Applicable.

D.4 : CONSTRUCTION OF A ML SOIL PROPERTY IMPUTER MODEL BASED ON KNN

To address gaps in our dataset for soil properties, needed to predict PFAS sorption, we used data from SoilGrids250m v2.0 (https://soilgrids.org/; Poggio et al., 2021) to develop a K-nearest neighbor (KNN) imputation model (see A.3.2). The number of soils considered was 2,039, with information on soil pH, C_{org}, CEC, and soil texture for each sampling coordinate. The soils were distributed worldwide, with a higher number of data entries for USA, South America, and African soils (Figure D.1).

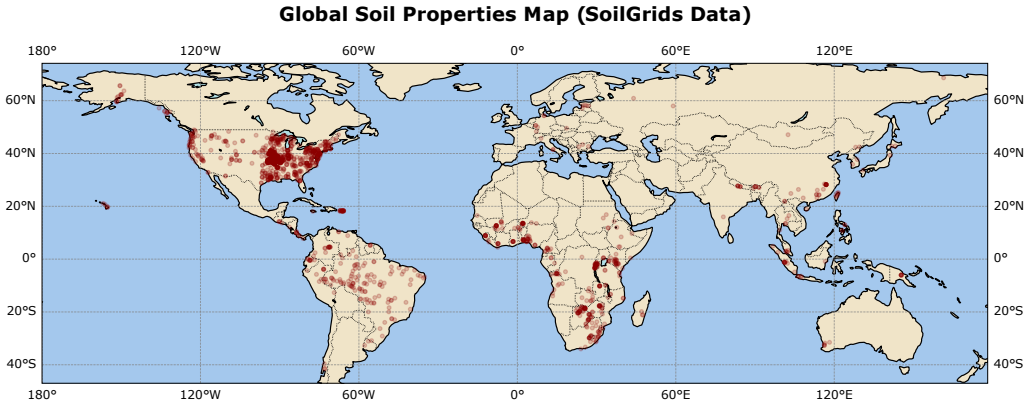


Figure D.1: Distribution of the 2,039 soil sampling coordinates across the globe resulting from the SoilGrids dataset.

The KNN imputer model allowed us to predict certain soil physicochemical properties that were lacking in the K_d (PFAS) dataset, especially soil pH (≈ 4% of the total data was imputed), CEC (≈ 20% of the total data was imputed), and sand (≈ 8% of the total data was imputed), silt (≈ 8% of the total data was imputed), and clay (≈ 8% of the total data was imputed) information. C_{org} was available for all accepted entries in Supplementary File 1 (download it from https://hydrogeochem.geo.uni-tuebingen.de/pfas).

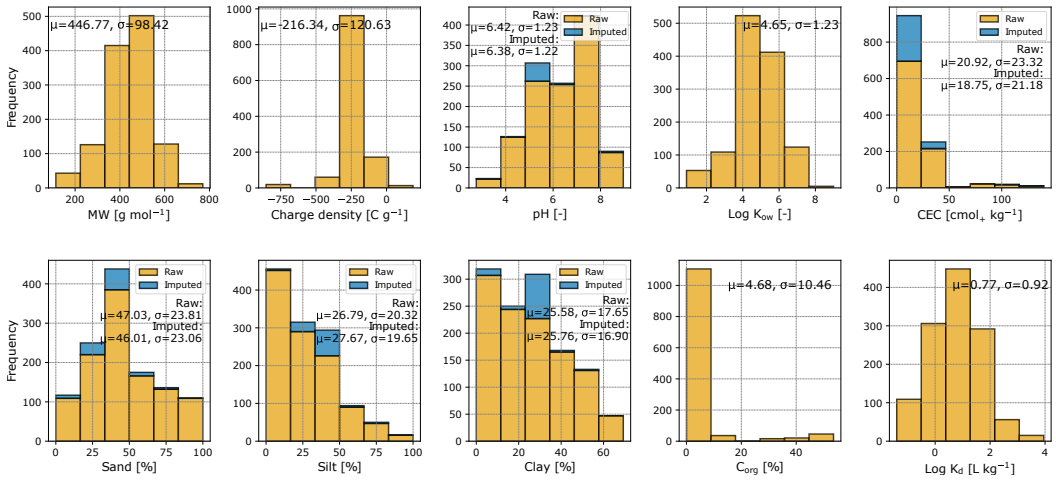


Figure D.2: Imputed soil- and PFAS-specific property ranges used for the model construction. The histograms illustrate the distributions of various soil and PFAS properties, comparing raw (orange) and imputed (blue) data. Statistical parameters (i.e., mean (μ) and standard deviation (σ) of the populations) are displayed.

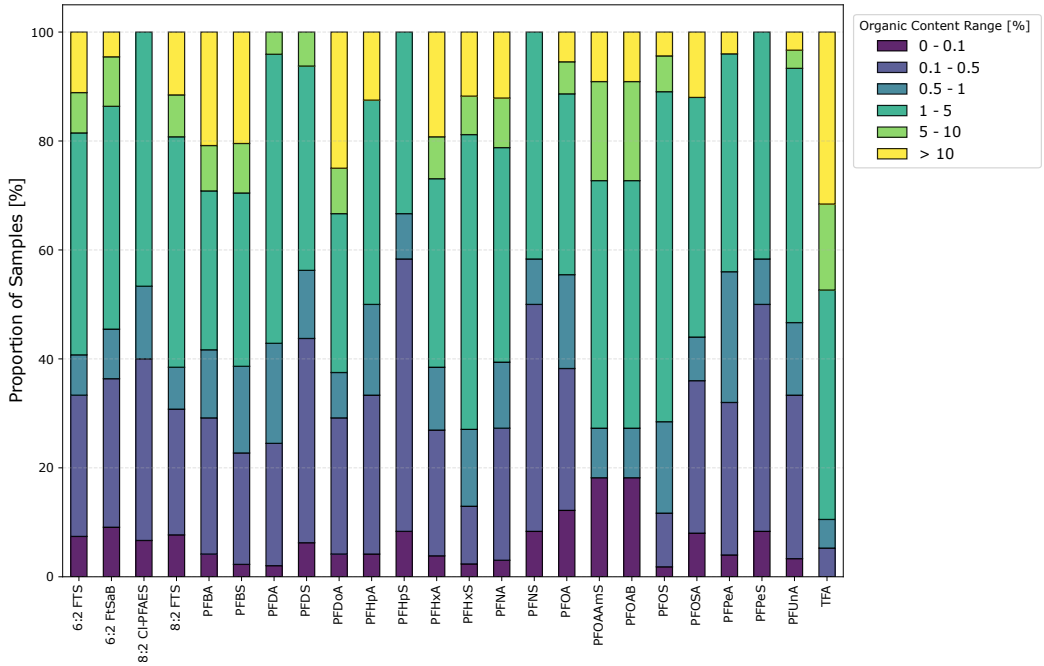


Figure D.3: Distribution of PFAS entries for several organic carbon content (C_{org}) ranges. The stacked bar plot shows the proportion of soil samples categorized into six C_{org}: 0 – 0.1%; 0.1 – 0.5%; 0.5 – 1%; 1 – 5%; 5 – 10%; >10%. The height of each bar represents the proportion of samples within each C_{org} range.

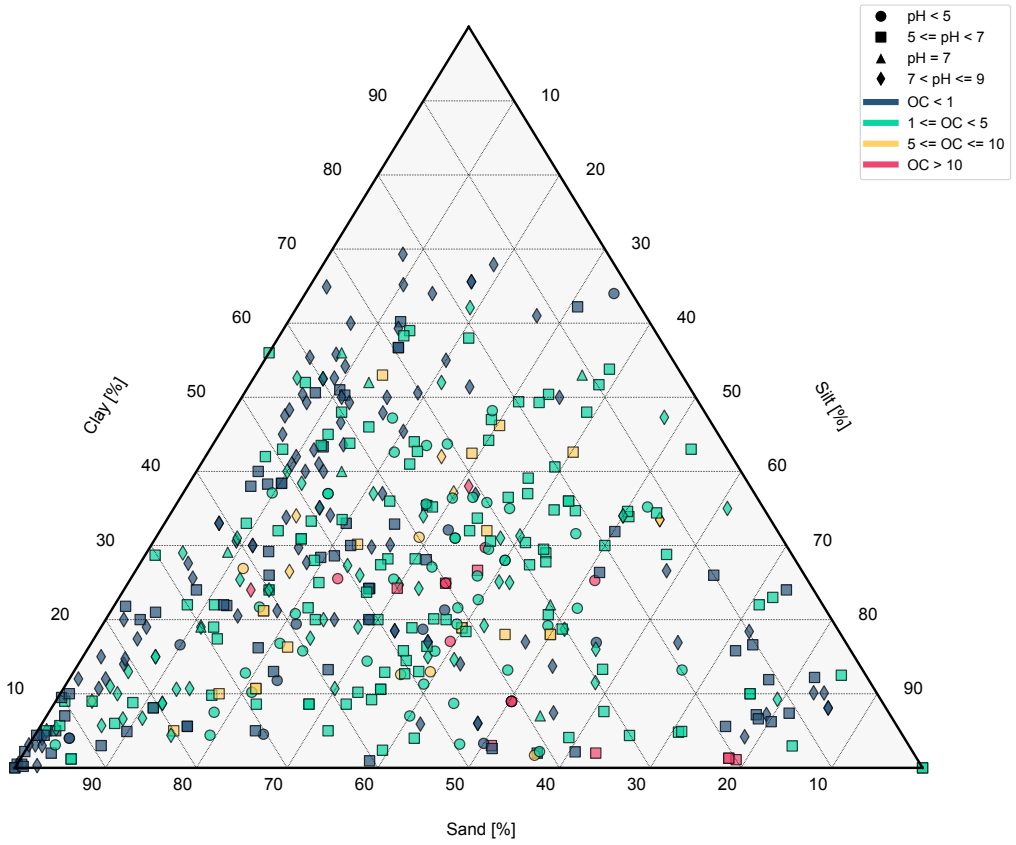


Figure D.4: Soil texture visualization illustrating the distribution of soil samples based on relative sand, silt, and clay contents, with overlaid data points representing different pH (marker shape, •: pH < 5; ■: $5 \leq$ pH < 7; ▲: pH $= 7$; ♦: $7 <$ pH ≤ 9) and C_{org} (%, marker color, blue: $C_{org} < 1\%$; green: $1 \leq C_{org} < 5\%$; yellow: $5 \leq C_{org} \leq 10\%$; red: $C_{org} > 10\%$) levels.

D.5 : STATISTICAL TESTS USED TO ASSESS LOG K_{OC} DISTRIBUTIONS ACROSS DIFFERENT PFAS SUBFAMILIES

A comprehensive statistical analysis was conducted to evaluate differences in log K_{OC} populations across various PFAS species, either by assessing the effect of the number of fluorinated carbons or the PFAS functionality. The statistical workflow was designed to ensure robust and precise comparisons, adhering to assumptions of normality and variance homogeneity wherever applicable. The flowchart of the statistical test selection process prior to assessing the log K_{OC} distributions is depicted in Figure D.5.

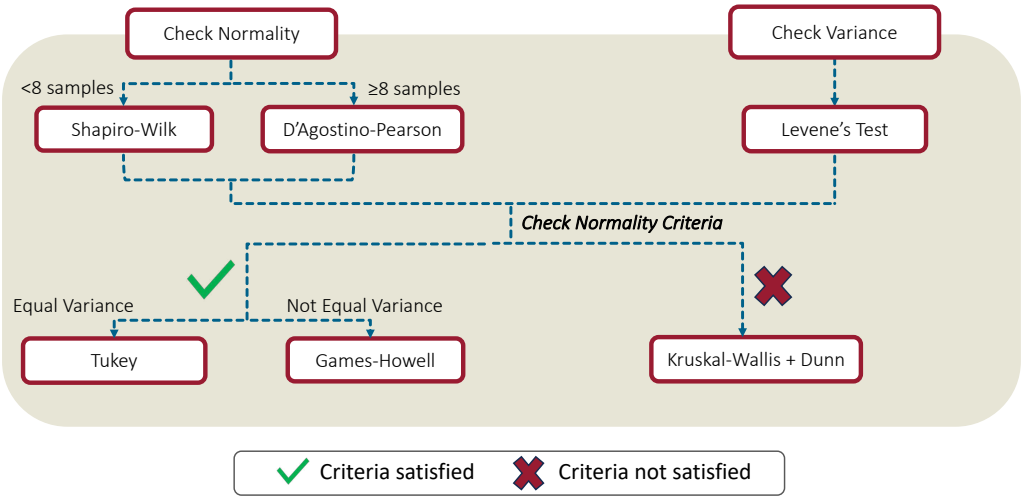


Figure D.5: Flowchart of the statistical test selection process prior to assessing the log K_{OC} distributions.

Normality Assessment

The distribution of log K_{OC} values within each group was evaluated depending on the number of data entries (n). The Shapiro-Wilk (Shapiro & Wilk, 1965) test (conducted at $\alpha = 0.05$) was applied to the log K_{OC} populations with $n < 8$ samples due to its reliability for small datasets. Log K_{OC} populations with $n \geq 8$ were assessed using the D'Agostino-Pearson test (D'Agostino & Pearson, 1973) (conducted at $\alpha = 0.05$). Groups were deemed normally distributed when $p > 0.05$.

Homogeneity of Variances

To determine whether the variances among log K_{OC} populations were comparable, Levene's test (Zimmerman, 2004) was applied (conducted at $\alpha = 0.05$). A $p > 0.05$ outcome from this test indicated equal variances across groups.

Statistical Testing Framework

Based on the results of both the normality and variance tests, appropriate statistical methods were selected. A one-way analysis of variance (ANOVA) combined with Tukey's honestly significant difference (HSD) (Tukey, 1949) was conducted at $\alpha = 0.05$ and applied

when all groups were normally distributed and had equal variances. The Welch's ANOVA (Welch, 1951) combined with the Games-Howell test (Games & Howell, 1976) was conducted at $\alpha = 0.05$ and applied when all groups were normally distributed but had unequal variances. The Kruskal-Wallis test (McKight & Najab, 2010) combined with Dunn's Test (Dinno, 2015) was conducted at $\alpha = 0.05$ and applied when any group deviated from normality. In all cases, groups with $p \geq 0.05$ were considered statistically indistinguishable and were assigned the same alphabetical label.

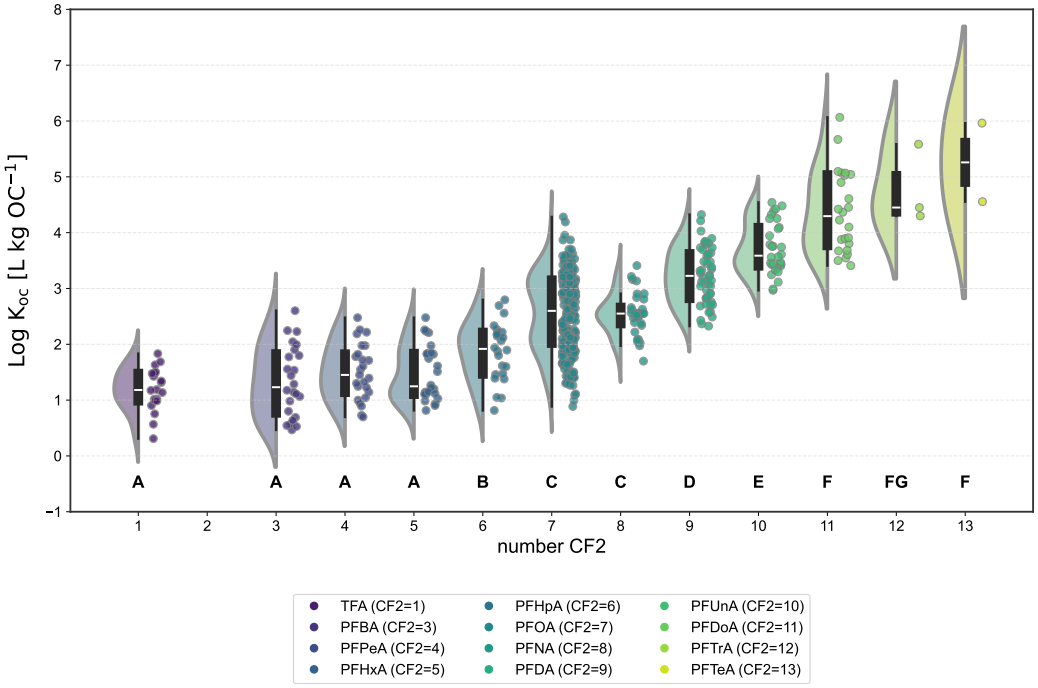


Figure D.6: Effect of PFAS chain length on log K_{OC} for the different PFCA.

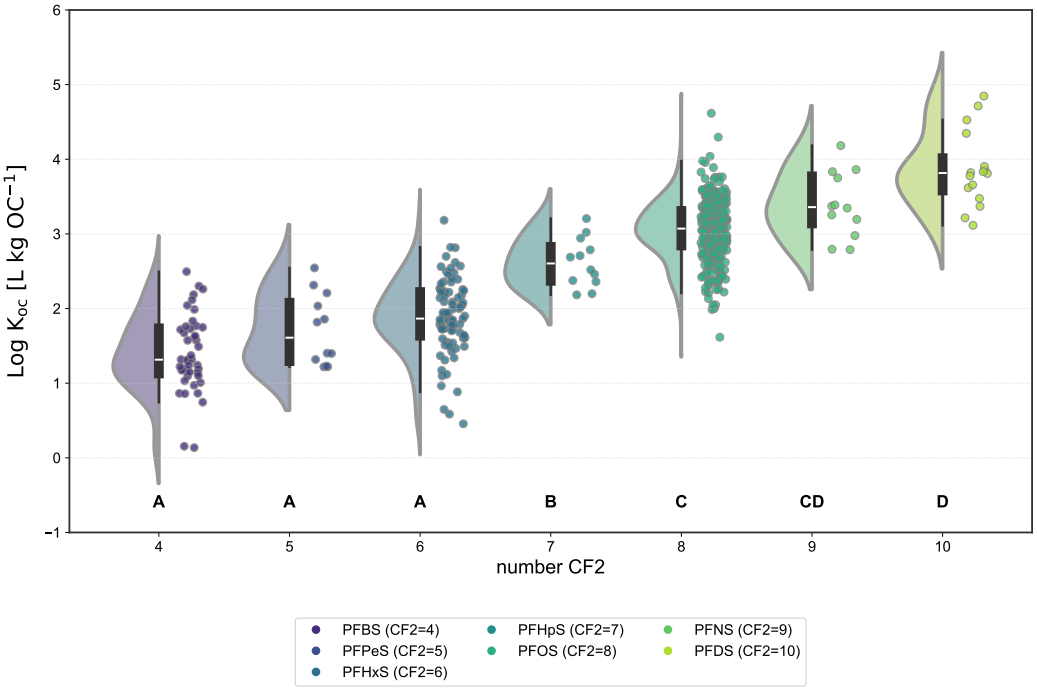


Figure D.7: Effect of PFAS chain length on log K_{OC} for the different PFSA.

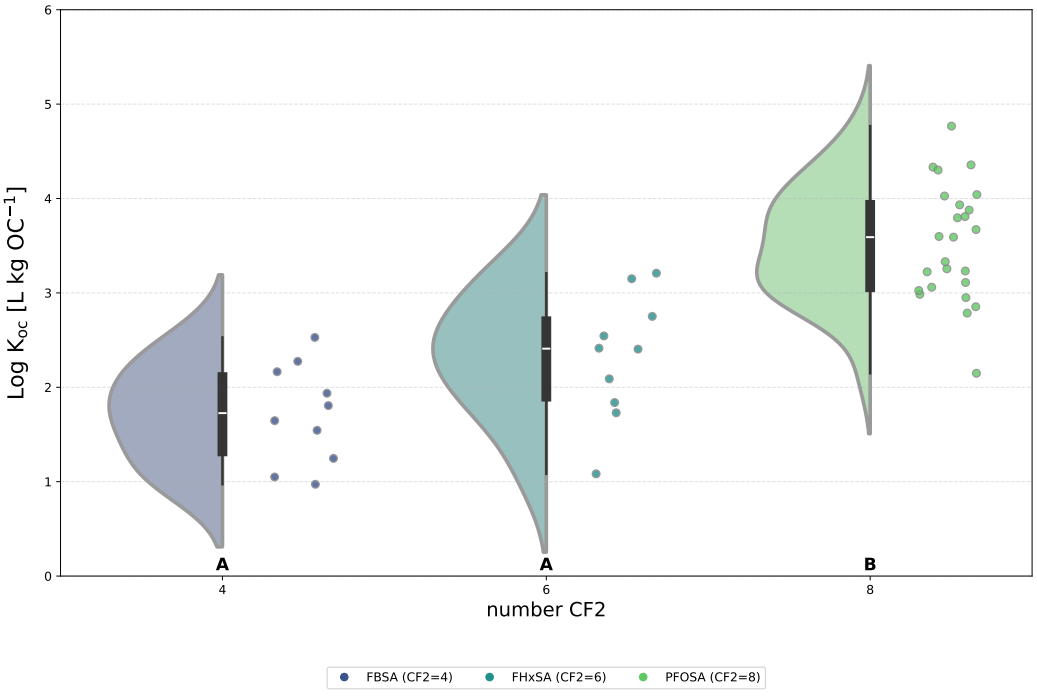


Figure D.8: Effect of PFAS chain length on log K_{OC} for the different FOSA.

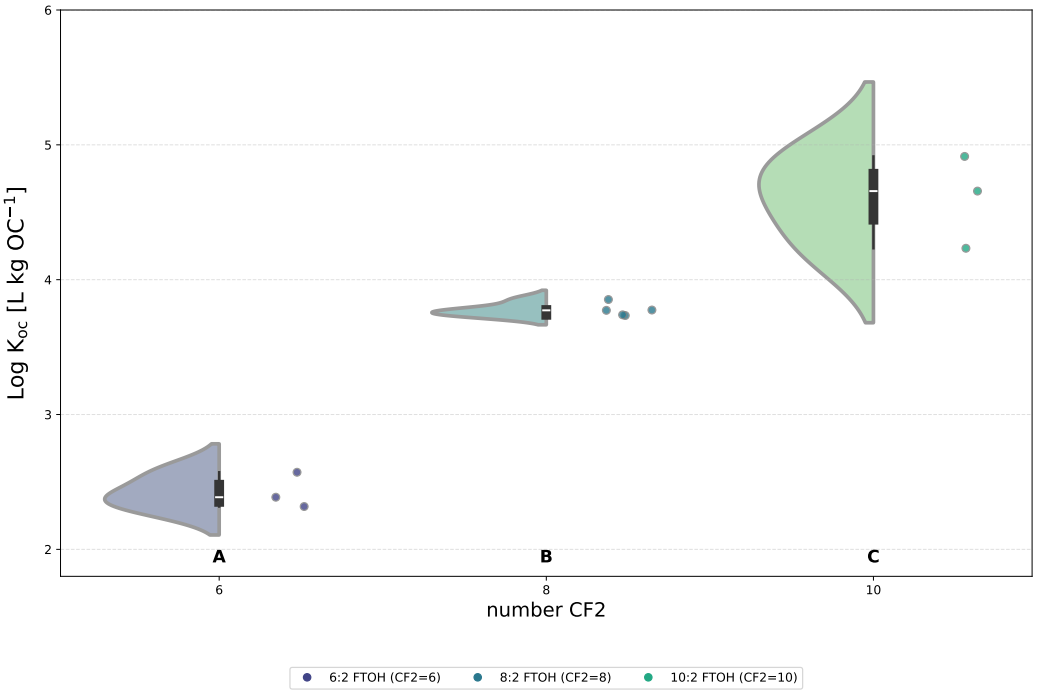


Figure D.9: Effect of PFAS chain length on log K_{OC} for the different FTOH.

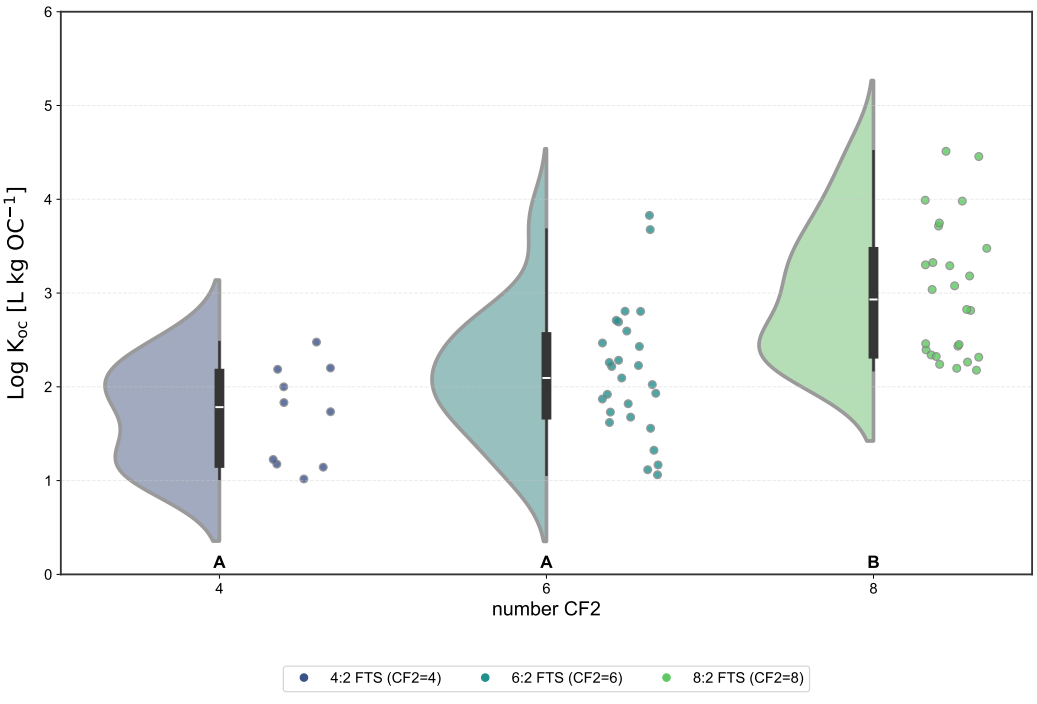


Figure D.10: Effect of PFAS chain length on log K_{OC} for the different FTS.

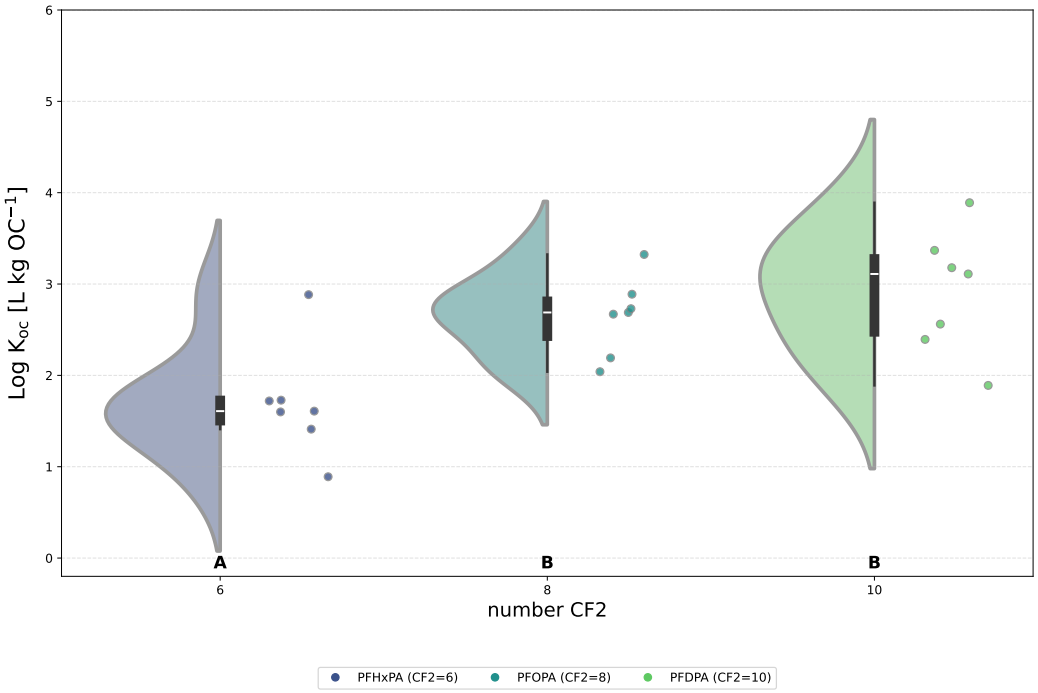


Figure D.11: Effect of PFAS chain length on log K_{OC} for the different PFPA.

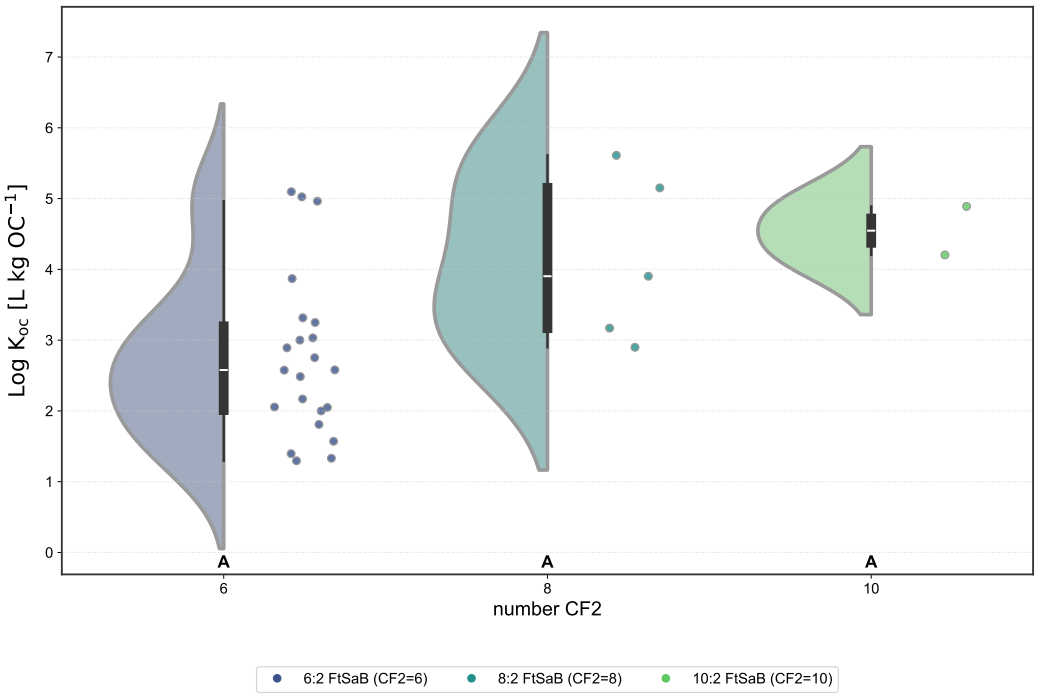


Figure D.12: Effect of PFAS chain length on log K_{OC} for the different Betaine.

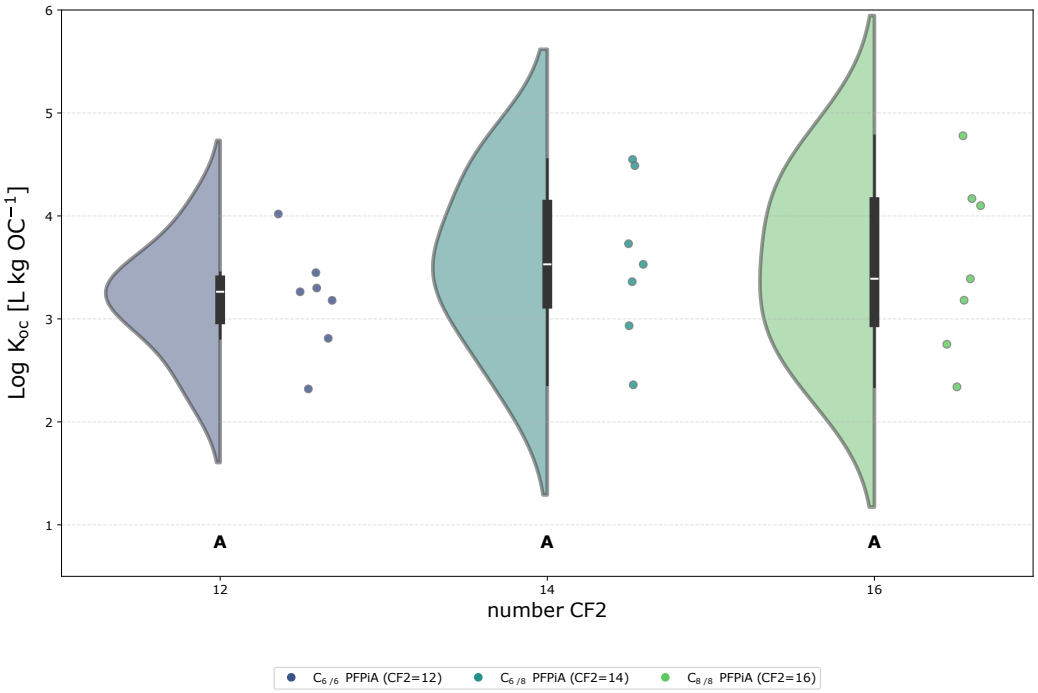


Figure D.13: Effect of PFAS chain length on log K_{OC} for the different PFPiA.

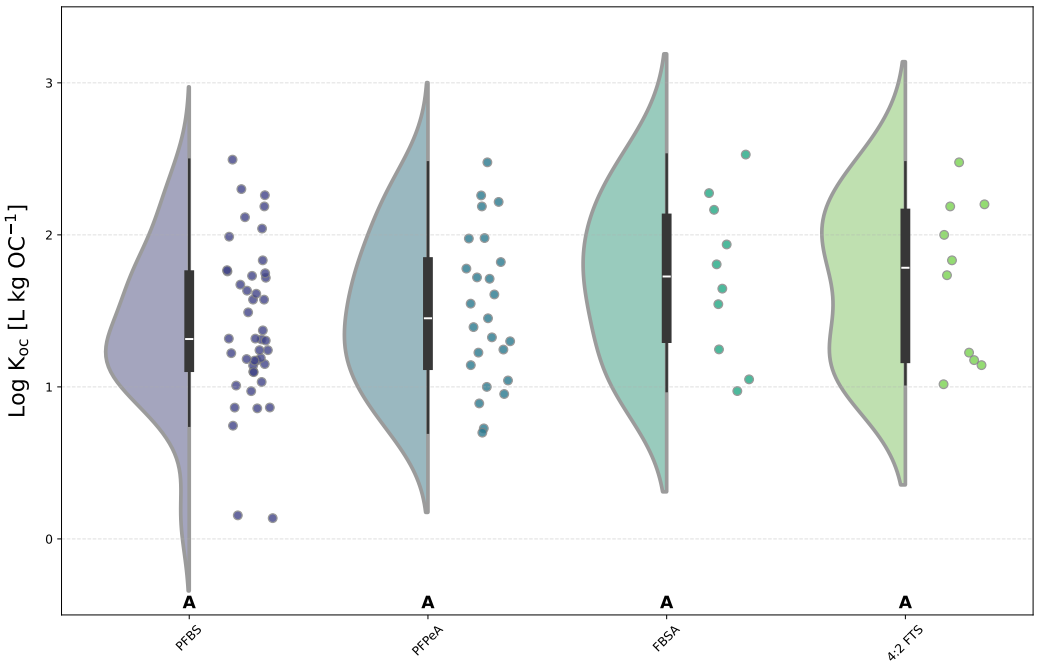


Figure D.14: Effect of PFAS functional group for a chain length of 4 fluorinated carbons.

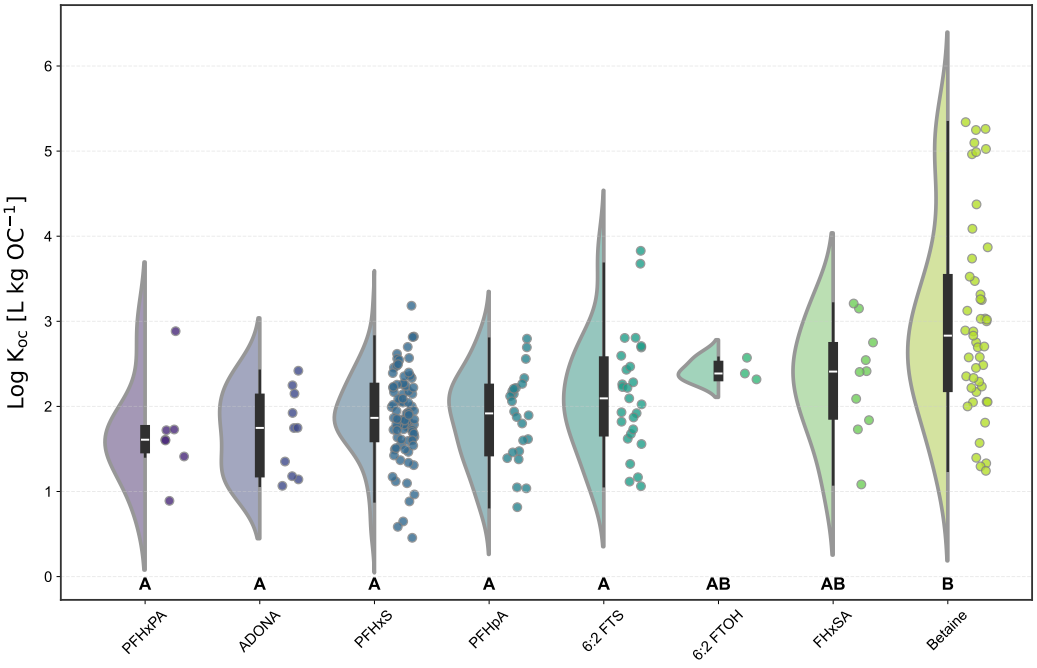


Figure D.15: Effect of PFAS functional group for a chain length of 6 fluorinated carbons.

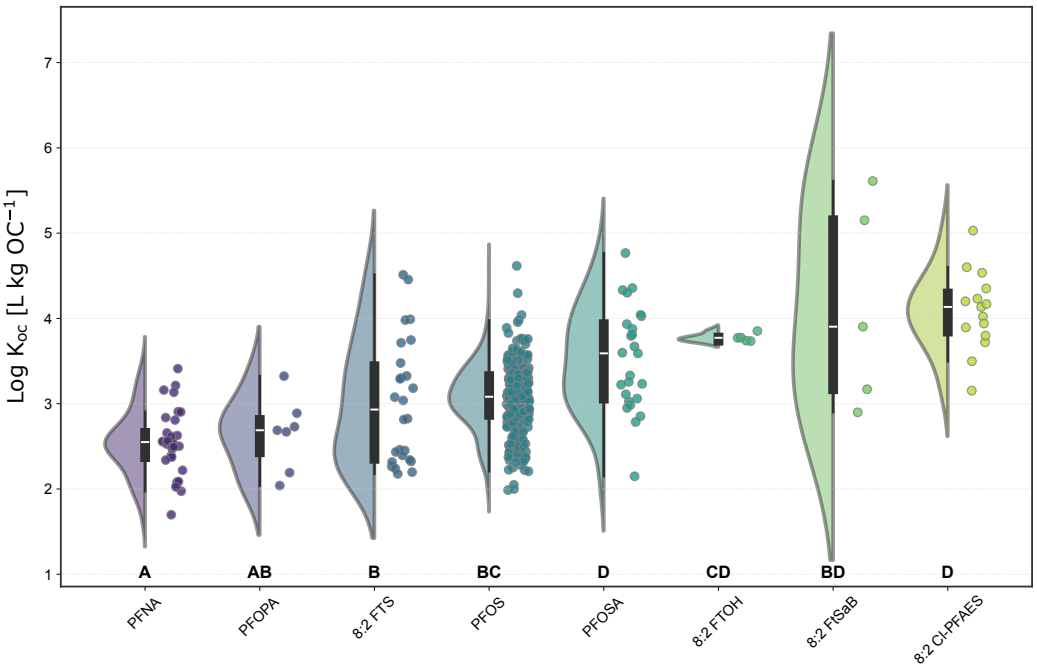


Figure D.16: Effect of PFAS functional group for a chain length of 8 fluorinated carbons.

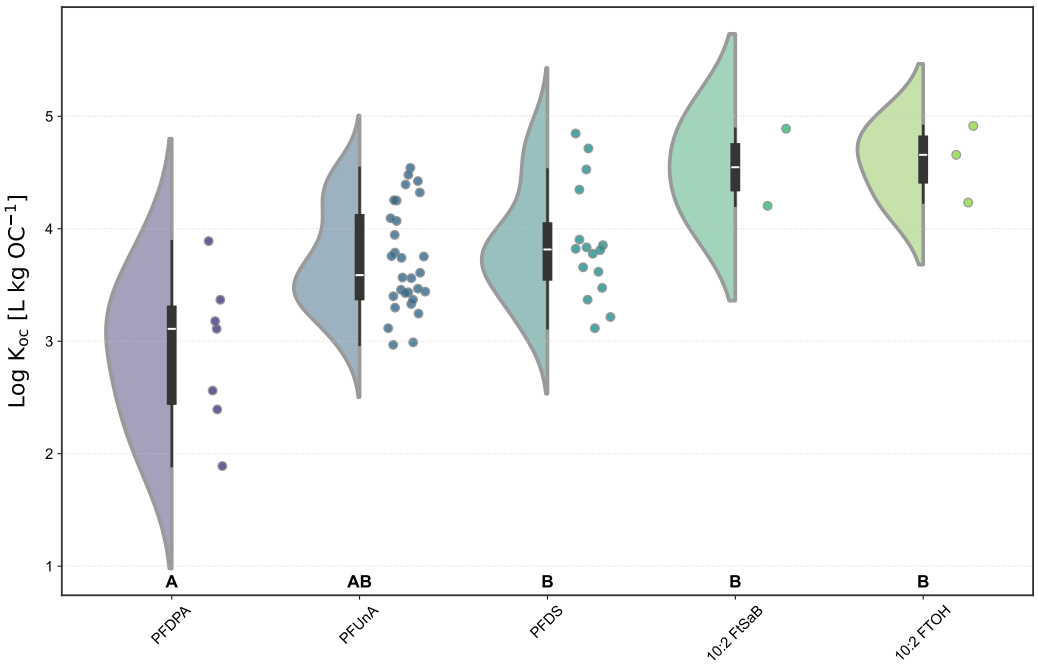


Figure D.17: Effect of PFAS functional group for a chain length of 10 fluorinated carbons.

D.6 : COMPARISON OF MODEL PERFORMANCE WITH OTHER AVAILABLE TOOLS

Our model's predictive metrics were compared against other available K_d (PFAS) predictive tools. EPISuite allows the derivation of K_{OC} values (estimated based on K_{OW} using KOWWIN v1.68) with the only input requirement being the CAS number and SMILES formula (Card et al., 2017). Therefore, it is a suitable first estimate to calculate K_d for any soil/PFAS pair as:

$$K_d = K_{OC} \cdot f_{OC}, \tag{D.5}$$

where K_{OC} [L kg OC^{-1}] is the result of EPISuite prediction for the tested PFAS, and f_{OC} is the fraction of soil organic carbon [kg OC kg soil^{-1}].

The thermodynamically-based mechanistic model presented by Higgins and Luthy (2007) was developed to predict the sorption of PFCA and PFSA in five sediments with $f_{OC} \leq 10\%$ via both electrostatic and hydrophobic interactions with the organic matter (Higgins & Luthy, 2007), as in Equation D.6:

$$K_d = f_{OC} \cdot \frac{F_{access} \cdot \bar{V}_W}{\rho_{OC} \cdot \bar{V}_i} \cdot \exp(-1) \cdot \exp\left(- \frac{\Delta G_{hyd,i} - z_i \cdot F \cdot \Psi_D}{(R \cdot T)} \right), \tag{D.6}$$

where f_{OC} is the fraction of soil organic carbon [kg OC kg soil^{-1}], ρ_{OC} is the organic matter density [kg OC L OC^{-1}], F_{access} [-] is the fraction of organic matter accessible for the sorbate i ($F_{access} \approx 0.34$), $\bar{V}_W$ and $\bar{V}_i$ are the molar volumes of water (0.018 L mol^{-1}) and sorbate, respectively, $\Delta G_{hyd,i}$ [KJ mol^{-1}] is the hydrophobic free energy change in sorption for the sorbate i, z_i [C] is the charge of the sorbate, F is the Faraday constant [C mol^{-1}], Ψ_D is the organic matter's electrostatic potential [V], R is the universal gas constant [J K^{-1} mol^{-1}], and T is temperature [K].

Fabregat-Palau and coworkers (Fabregat-Palau et al., 2021) developed a model applicable to any soil type to predict K_d values for PFCA and PFSA with a number of fluorinated carbons ranging 3–11 based on independent interactions of PFAS with soil C_{org} and mineral (i.e., silt and clay fractions) domains, as in Equation D.7:

$$K_d = K_{OC} \cdot f_{OC} + K_{MIN} \cdot f_{MIN} = 10^{(0.47 \cdot CF_2 - 0.70)} \cdot f_{OC} + 10^{(0.32 \cdot CF_2 - 1.70)} \cdot f_{S+C}, \tag{D.7}$$

where K_{MIN} is the mineral-normalized sorption coefficient [L kg reactive mineral fraction^{-1}] and f_{S+C} is the fraction of reactive mineral fraction [kg silt+clay kg soil^{-1}].

Knight and coworkers (Knight et al., 2019) developed a K_d (PFOA) model after running a multiple linear regression to 100 soils with $f_{OC} \leq 3.5\%$, pH ranging 4.9–8.6, and silt+clay contents ranging 5–88%. These property ranges, therefore, set the model applicability range, which according to our data assessment had the following form:

$$K_{d,PFOA} = -0.52 \cdot pH + 3.2 \cdot C_{org} + 0.04 \cdot (silt + clay) + 5.75, \tag{D.8}$$

where pH [-], C_{org} [%], and silt + clay [%] are the soil physicochemical properties.

Umeh and coworkers (Umeh et al., 2021) developed a K_d (PFOS) model based on an artificial neural network trained on sorption data from 114 soils of contrasting properties but limited to soils with $f_{OC} \leq 10\%$. While their model did not contain an explicit equation, they developed an online platform to predict K_d (PFOS).

Similarly, Xie and coworkers (Xie et al., 2024) recently developed a K_d (PFAS) prediction model based on the application of random forest machine learning approaches to a K_d (PFAS) dataset covering a wide range of PFAS (including PFCA, PFSA, FTS, FOSA, and other novel PFAS) and soil properties. Unfortunately, their model did not contain an explicit equation, nor provided a platform to test the model.

Table D.2 summarizes some of the reported metrics in each study, where we additionally include the prediction of our model and additional prediction metrics originating from EPISuite (Equation D.5):

Model	Type of Validation	Number of Validation Data	RPD	Explained Variance (%)
EPISuite (This study)	Independent test set (20% of the total)	231	1.33	44
Higgins and Luthy, 2006	Independent test set (100% of the total)	≈ 80	N.A.	93
Fabregat-Palau et al., 2021	Independent test set (33% of the total)	121	1.88	76
Knight et al., 2019	95% of their own training set	95	1.83	62
Umeh et al., 2021	Cross-validation of their own dataset	114	N.A.	84
Xie et al., 2024	Cross-validation of their own dataset	2,328	N.A.	93
PSSM * (This study)	Independent test set (20% of the total)	231	3.13	89

Note: N.A. Not Available; *:PFAS Sorption Stacking Model

Table D.2: Reported metrics evaluating the performance for different K_d (PFAS) prediction tools.

Based on the metrics listed in TableD.2, our model outperformed the prediction quality of other currently available tools, as exemplified by the highest RPD value across studies

and with a higher explained variability. Furthermore, the good prediction of K_d (PFAS) data originating from various studies and obtained under various batch conditions (i.e., differing nature of the contact solution) over those studies assessing sorption for specific PFAS under certain constant batch conditions (e.g., Knight et al., 2019 and Umeh et al., 2021) highlights the broader applicability of our model.

Figure D.18 exemplifies the prediction of our test set for three of the tools listed in Table D.2: EPISuite, Fabregat-Palau et al., 2021, and our developed PSSM. EPISuite and PSSM predictions were examined across the overall test set, while predictions from Fabregat-Palau et al., 2021, were only applied to the test set for those PFCA and PFSA species with fluorinated carbon numbers ranging from 3 to 11.

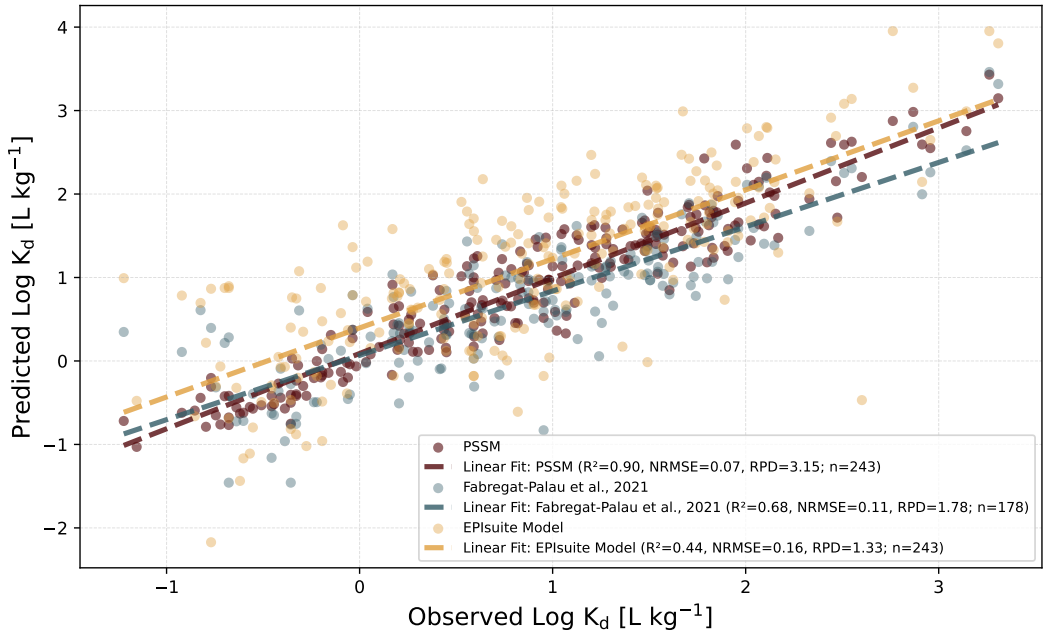


Figure D.18: Assessment of different models (i.e., EPISuite, Fabregat-Palau et al., 2021, and PSSM) based on our test set.

As displayed in Figure D.18, the prediction accuracy originating from EPISuite was the lowest (i.e., 44% of the data variability explained, RPD = 1.3, $n = 243$, $p < 0.001$), likely due to K_d (PFAS) output relying only on K_{OC} and f_{OC} values. Despite being useful for an early estimation, EPISuite output does not consider the sorption in reactive mineral fractions, which may be relevant for soils with low C_{org} and especially for compounds such as short-chain PFCA and cationic and zwitterionic PFAS (Fabregat-Palau et al., 2021, 2024).

Fabregat-Palau et al., 2021 addressed this issue by including these reactive mineral (i.e., silt and clay) fractions, thus achieving a better prediction performance (i.e., 68% of the data variability explained, RPD = 1.7, $n = 178$, $p < 0.001$) for the PFCA and PFSA species with fluorinated carbon numbers ranging from 3 to 11. Nonetheless, the developed model, which was built upon linear relationships between K_{OC} and K_{MIN} with the number of fluorinated carbons of the PFAS, did not include the effect of other properties (e.g., pH) in sorption or potential non-linear relationship among variables.

The PSSM model addresses this by upgrading the pool of PFAS and including other soil properties beyond C_{org} and reactive mineral fractions, which are modeled through

non-linear approaches typical in ML, allowing the prediction performance of the overall dataset to rise (i.e., 90% of the data variability explained, RPD = 3.2, $n = 243$, $p < 0.001$).

D.7 : ADDITIONAL GEOSPATIAL K_d (PFAS) MAPS

The PSSM model developed in this study was applied to each sampling location of the topsoil LUCAS 2009 repository data for European soils (European Soil Data Centre (ESDAC), https://esdac.jrc.ec.europa.eu/content/lucas-2009-topsoil-data) that had available information for all the model input parameters required (i.e., pH, C_{org}, CEC, sand, silt, and clay contents). The total number of soil data points was 21,904.

The soil characteristics of the dataset were preliminarily screened (Figure D.19). C_{org} was generally low ($\leq 5\%$) across southern Europe but increased ($\geq 20\%$) in some regions of northern Europe and the Scandinavian Peninsula. CEC showed a similar pattern to C_{org}, with values generally spanning 10–40 $cmol_+$ kg^{-1}. Soil pH across central and southern Europe showed slightly alkaline conditions (i.e., pH ≈ 7–9), but had acidic conditions (i.e., pH ≈ 4–6) in northern Europe, Portugal, and the Scandinavian Peninsula. Regarding soil textural information, central European soils displayed higher silt contents (i.e., ≈40–80%). Clay contents were generally low (i.e., <40%) but increased (i.e., ≈50–80%) in Southeast Europe. Sand contents generally ranged 10–40% in central Europe but increased to contents >60% in countries like Denmark, northern Germany, and Poland.

The developed PSSM is able to produce geospatial sorption information for PFAS. As an example, we display the outcomes for four PFAS compounds (i.e., TFA (Figure D.20), PFOA (Figure D.21), PFOS (Figure D.22), and PFOSB (Figure D.23)), although the end-user is able to access the geospatial K_d information for all 47 PFAS included in the model by using PFASorptionML at https://hydrogeochem.geo.uni-tuebingen.de.

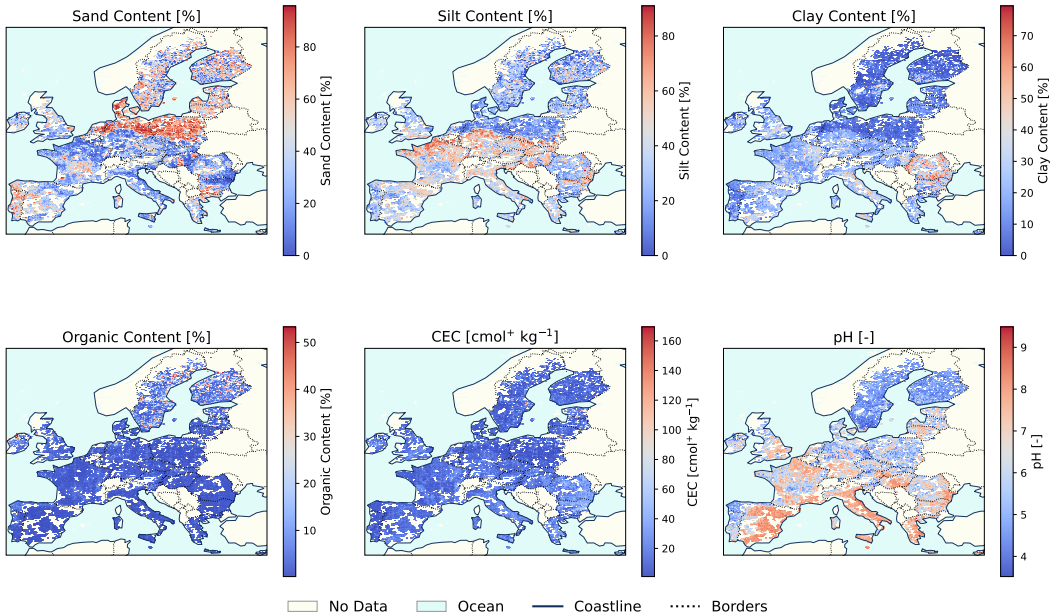


Figure D.19: Geospatial characteristics of the European soils.

As observed in Figure D.20, the K_d values for TFA were low (≈ 0.3 L kg^{-1}) across Europe, indicating very low retention in soil and high mobility to the groundwater table, although higher values (≈ 10 L kg^{-1}) are observed in some regions of the Scandinavian Peninsula, likely as a result of higher C_{org} and acidic soil characteristics. TFA is considered a very persistent and very mobile (vPvM) substance that has been increasing in concentration within diverse environmental media including rain and drinking water, with concentrations one order of magnitude higher than those of other PFAS (Arp et al., 2024). Our geospatial K_d (TFA) distribution demonstrates the vulnerability of European groundwater towards TFA leaching resulting from rain and application in topsoil.

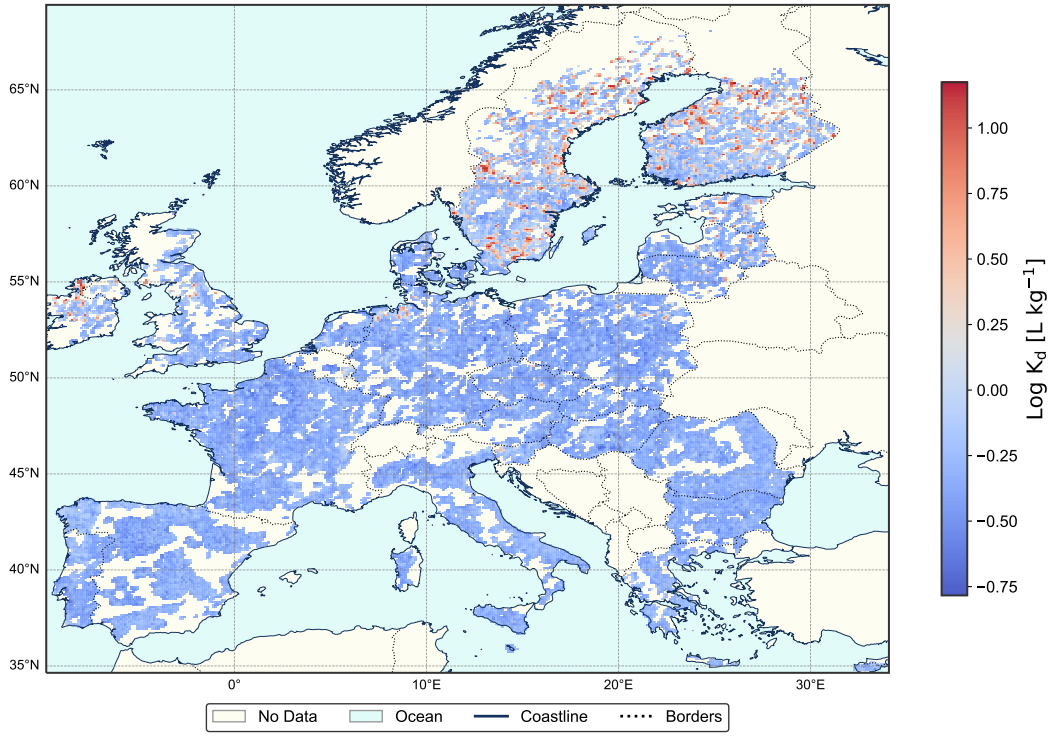


Figure D.20: Predicted log K_d values for TFA across Europe using the PSSM model based on soil properties from the LUCAS 2009 repository.

Regarding PFOA, K_d values spanning ≈3–30 L kg^{-1} may be anticipated across European soils according to its properties (Figure D.21). These values suggest a low to moderate sorption to soil particles and, therefore, differing mobility in seepage water. Of relevance are some regions in northern France, Belgium, Denmark, the Netherlands, and north-west Germany, which display relatively low K_d (PFOA) (≈ 3 L kg^{-1}), and therefore lower retardation to the groundwater table. Higher PFAS concentrations are expected in topsoil in these regions due to the presence of multiple PFAS hotspots (Moghadasi et al., 2023).

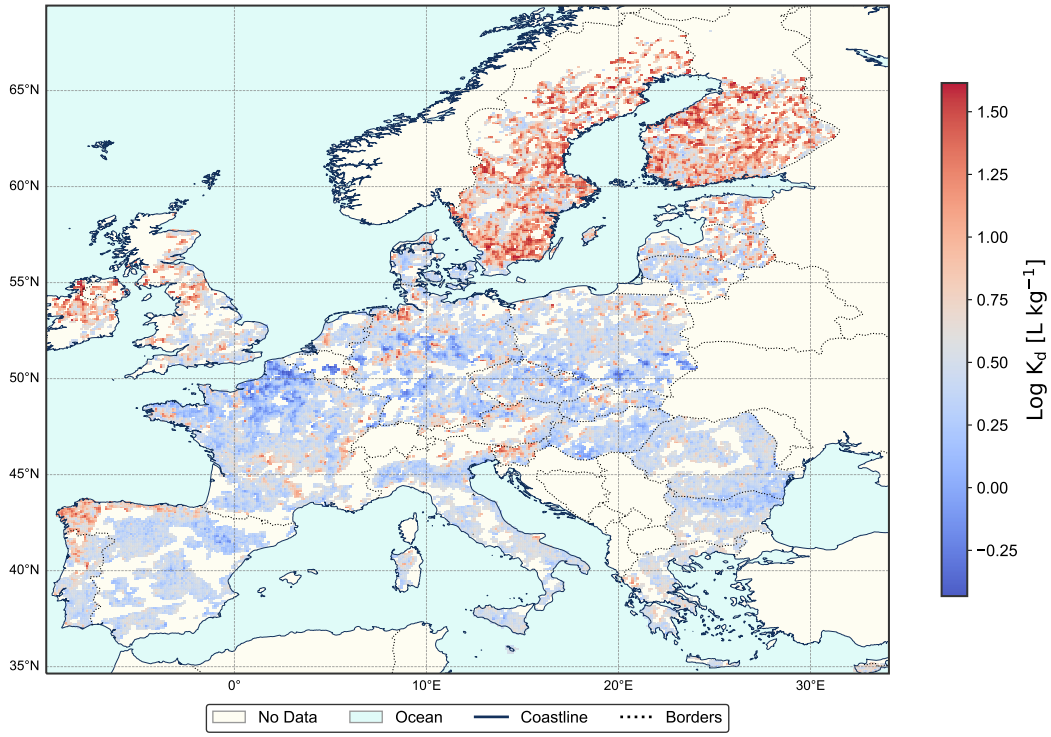


Figure D.21: Predicted log K_d values for PFOA across Europe using the PSSM model based on soil properties from the LUCAS 2009 repository.

Regarding PFOS, K_d values ranging from ≈10 to ≈200 L kg^{-1} are anticipated across European soils (Figure D.22), indicating higher retardation to the groundwater table than that expected for PFOA due to stronger sorption to soil particles. Geospatial K_d (PFOS) predictions agree with those observed for PFOA, highlighting a potential threat to groundwater from northern France, Belgium, Denmark, Netherlands, and northwest Germany, where higher PFAS concentrations are expected in topsoil due to the presence of multiple PFAS hotspots (Moghadasi et al., 2023).

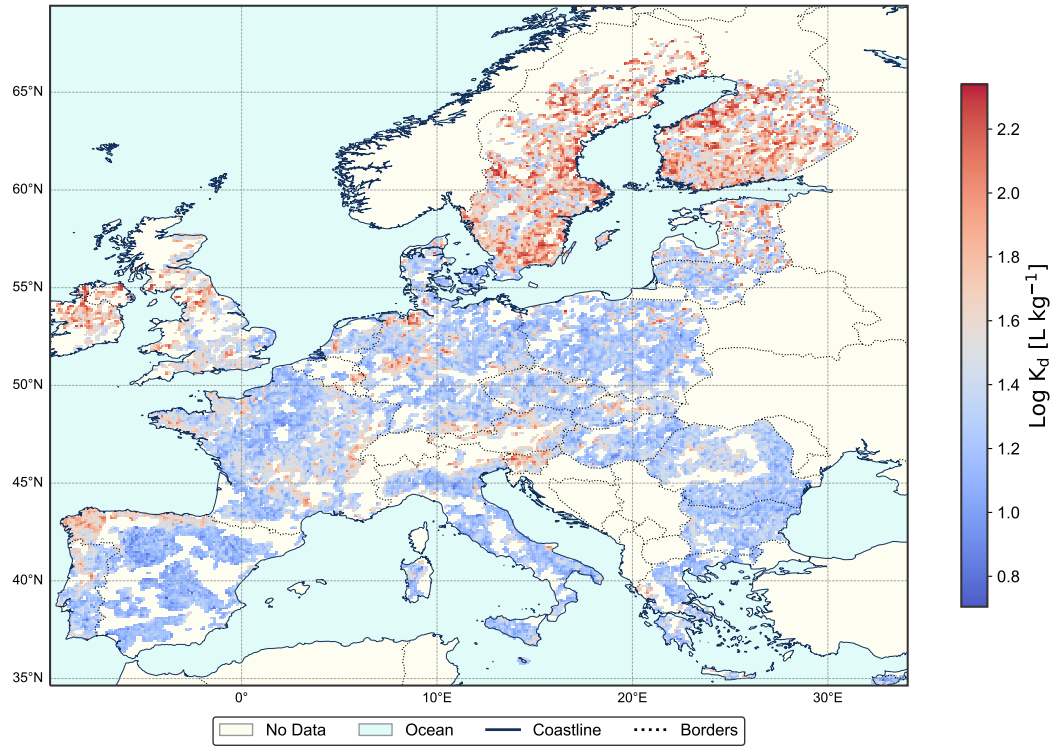


Figure D.22: Predicted log K_d values for PFOS across Europe using the PSSM model based on soil properties from the LUCAS 2009 repository.

Regarding PFOSB, selected here as a representative Betaine compound, predicted K_d values across European soils span ≈2 to 200 L kg^{-1} (Figure D.23). Sorption geospatial distribution differs from that observed for TFA, PFOA, and PFOS, with higher values observed in Central Europe and the Scandinavian Peninsula. PFOSB, as well as other Betaine-like PFAS, composes the majority of the PFAS burden at AFFF-impacted sites Schüßler et al., 2024. The higher sorption of PFOSB in locations of Central Europe, which generally have soils with slight alkalinity and relatively higher abundances of silt and clay fractions (see Figure D.19), may result from the electrostatic interaction between the cationic group of PFOSB species and the negatively charged clay surfaces (Barzen-Hanson et al., 2017). On the other hand, the higher sorption observed in the Scandinavian Peninsula may result from the higher amount of C_{org} in the soil, with an implicit higher CEC and therefore a higher number of sorption sites able to interact with PFOSB through cation exchange mechanisms (Barzen-Hanson et al., 2017; Fabregat-Palau et al., 2024).

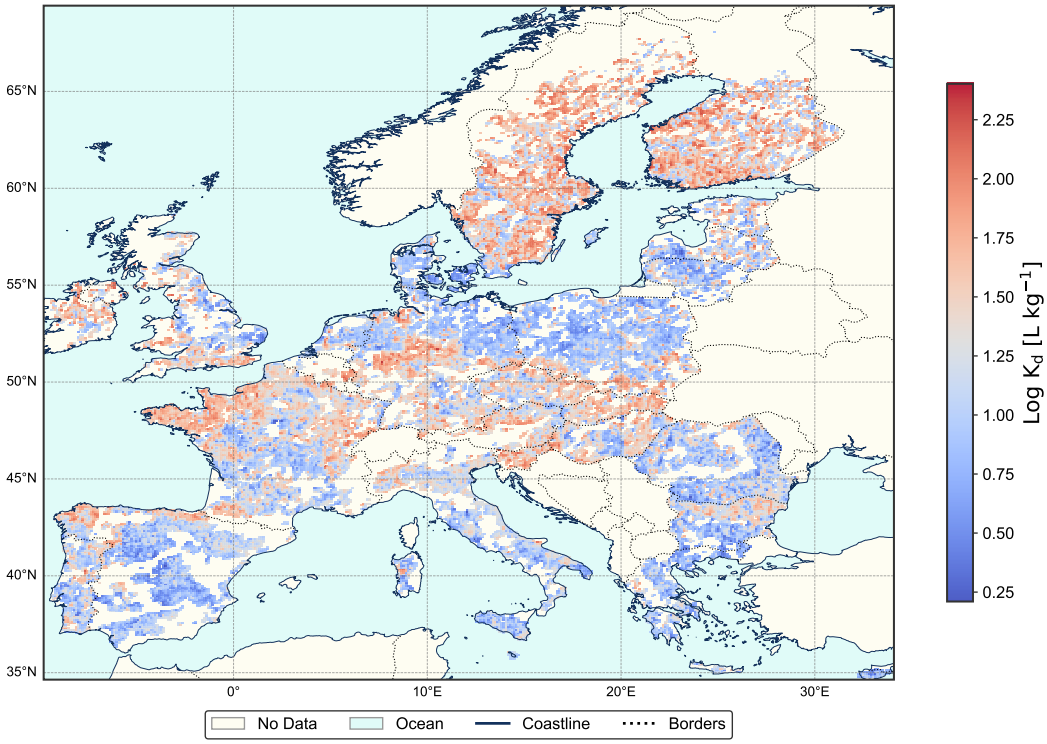


Figure D.23: Predicted log K_d values for PFOSB across Europe using the PSSM model based on soil properties from the LUCAS 2009 repository.

BIBLIOGRAPHY

Alengebawy, A., Abdelkhalek, S. T., Qureshi, S. R., & Wang, M.-Q. (2021). Heavy Metals and Pesticides Toxicity in Agricultural Soil and Plants: Ecological Risks and Human Health Implications. *Toxics*, *9*, 42. doi: 10.3390/toxics9030042.

Allen-King, R. M., Grathwohl, P., & Ball, W. P. (2002). New modeling paradigms for the sorption of hydrophobic organic chemicals to heterogeneous carbonaceous matter in soils, sediments, and rocks. *Advances in Water Resources*, *25*, 985–1016. doi: 10.1016/S0309-1708(02)00045-3.

Allgaier, J., & Pryss, R. (2024). Cross-Validation Visualized: A Narrative Guide to Advanced Methods. *Machine Learning and Knowledge Extraction*, *6*, 1378–1388. doi: 10.3390/make6020065.

Allgeier, J., & Cirpka, O. A. (2023). Surrogate-Model Assisted Plausibility-Check, Calibration, and Posterior-Distribution Evaluation of Subsurface-Flow Models. *Water Resources Research*, *59*, e2023WR034453. doi: 10.1029/2023WR034453.

Arp, H. P. H., Gredelj, A., Glüge, J., Scheringer, M., & Cousins, I. T. (2024). The Global Threat from the Irreversible Accumulation of Trifluoroacetic Acid (TFA). *Environmental Science & Technology*, *58*, 19925–19935. doi: 10.1021/acs.est.4c06189.

Asadollahi, M., Stumpp, C., Rinaldo, A., & Benettin, P. (2020). Transport and Water Age Dynamics in Soils: A Comparative Study of Spatially Integrated and Spatially Explicit Models. *Water Resources Research*, *56*, e2019WR025539. doi: 10.1029/2019WR025539.

Babaei, M., Alkhatib, A., & Pan, I. (2015). Robust optimization of subsurface flow using polynomial chaos and response surface surrogates. *Computational Geosciences*, *19*, 979–998. doi: 10.1007/s10596-015-9516-5.

Bălan, S. A., Mathrani, V. C., Guo, D. F., & Algazi, A. M. (2021). Regulating PFAS as a Chemical Class under the California Safer Consumer Products Program. *Environmental Health Perspectives*, *129*, 025001. doi: 10.1289/EHP7431.

Bandow, N., Finkel, M., Grathwohl, P., & Kalbe, U. (2019). Influence of flow rate and particle size on local equilibrium in column percolation tests using crushed masonry. *Journal of Material Cycles and Waste Management*, *21*, 642–651. doi: 10.1007/s10163-019-00827-3.

Barceloux, D. G., & Barceloux, D. (1999). Vanadium. *Journal of Toxicology: Clinical Toxicology*, *37*, 265–278. doi: 10.1081/CLT-100102425.

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

Bayar, S., Demir, I., & Engin, G. O. (2009). Modeling leaching behavior of solidified wastes using back-propagation neural networks. *Ecotoxicology and Environmental Safety*, *72*, 843–850. doi: 10.1016/j.ecoenv.2007.10.019.

Bazoobandi, A., Emamgholizadeh, S., & Ghorbani, H. (2019). Estimating the amount of cadmium and lead in the polluted soil using artificial intelligence models. *European Journal of Environmental and Civil Engineering*, *26*, 933–951. doi: 10.1080/19648189.2019.1686429.

Belkina, A. C., Ciccolella, C. O., Anno, R., Halpert, R., Spidlen, J., & Snyder-Cappione, J. E. (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications*, *10*, 5415. doi: 10.1038/s41467-019-13055-y.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*, 281–305. http://jmlr.org/papers/v13/bergstra12a.html

Bernard, L., Maron, P. A., Mougel, C., Nowak, V., Lévêque, J., Marol, C., Balesdent, J., Gibiat, F., & Ranjard, L. (2009). Contamination of Soil by Copper Affects the Dynamics, Diversity, and Activity of Soil Bacterial Communities Involved in Wheat Decomposition and Carbon Storage. *Applied and Environmental Microbiology*, *75*, 7565–7569. doi: 10.1128/AEM.00616-09.

Berrar, D. (2019). Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542–545). Elsevier. doi: 10.1016/B978-0-12-809633-8.20349-X.

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*, 1039–1082. doi: 10.1007/s10994-017-5633-9.

BMUV. (2023). *Waste management in germany 2023* (tech. rep.). Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV). https://www.bmuv.de/fileadmin/Daten_BMU/Pools/Broschueren/abfallwirtschaft_2023_en_bf.pdf

Boving, T. B., & Grathwohl, P. (2001). Tracer diffusion coefficients in sedimentary rocks: Correlation to porosity and hydraulic conductivity. *Journal of Contaminant Hydrology*, *53*, 85–100. doi: 10.1016/S0169-7722(01)00138-3.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*, 49–64. doi: 10.1007/BF00117832.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. doi: 10.1023/A:1010933404324.

Brusseau, M. L. (2023). Influence of chain length on field-measured distributions of PFAS in soil and soil porewater. *Journal of Hazardous Materials Letters*, *4*, 100080. doi: 10.1016/j.hazl.2023.100080.

Brusseau, M. L., Anderson, R. H., & Guo, B. (2020). PFAS concentrations in soils: Background levels versus contaminated sites. *Science of The Total Environment*, *740*, 140017. doi: 10.1016/j.scitotenv.2020.140017.

Buck, R. C., Franklin, J., Berger, U., Conder, J. M., Cousins, I. T., De Voogt, P., Jensen, A. A., Kannan, K., Mabury, S. A., & Van Leeuwen, S. P. (2011). Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. *Integrated Environmental Assessment and Management*, *7*, 513–541. doi: 10.1002/ieam.258.

Bundesgesetzblatt. (2021). Verordnung zur Einführung einer Ersatzbaustoffverordnung, zur Neufassung der Bundes-Bodenschutz- und Altlastenverordnung und zur änderung der Deponieverordnung und der Gewerbeabfallverordnung. Bundesgesetzblatt Teil I Nr. 43, Bonn, 16. Juli 2021. https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&start=//*[@attr_id=%27bgbl121s2598.pdf%27]#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl121s2598.pdf%27%5D__1733485420960

Butera, S., Hyks, J., Christensen, T. H., & Astrup, T. F. (2015). Construction and demolition waste: Comparison of standard up-flow column and down-flow lysimeter leaching tests. *Waste Management*, *43*, 386–397. doi: 10.1016/j.wasman.2015.04.032.

Caglar Gencosman, B., & Eker Sanli, G. (2021). Prediction of Polycyclic Aromatic Hydrocarbons (PAHs) Removal from Wastewater Treatment Sludge Using Machine Learning Methods. *Water, Air, & Soil Pollution*, *232*, 87. doi: 10.1007/s11270-021-05049-8.

Cai, W., Navarro, D. A., Du, J., Ying, G., Yang, B., McLaughlin, M. J., & Kookana, R. S. (2022). Increasing ionic strength and valency of cations enhance sorption through hydrophobic interactions of PFAS with soil surfaces. *Science of The Total Environment*, *817*, 152975. doi: 10.1016/j.scitotenv.2022.152975.

Campos Pereira, H., Ullberg, M., Kleja, D. B., Gustafsson, J. P., & Ahrens, L. (2018). Sorption of perfluoroalkyl substances (PFASs) to an organic soil horizon – Effect of cation composition and pH. *Chemosphere*, *207*, 183–191. doi: 10.1016/j.chemosphere.2018.05.012.

Campos-Pereira, H., Makselon, J., Kleja, D. B., Prater, I., Kögel-Knabner, I., Ahrens, L., & Gustafsson, J. P. (2022). Binding of per- and polyfluoroalkyl substances (PFASs) by organic soil materials with different structural composition – Charge- and concentration-dependent sorption behavior. *Chemosphere*, *297*, 134167. doi: 10.1016/j.chemosphere.2022.134167.

Card, M. L., Gomez-Alvarez, V., Lee, W.-H., Lynch, D. G., Orentas, N. S., Lee, M. T., Wong, E. M., & Boethling, R. S. (2017). History of EPI Suite™ and future perspectives on chemical property estimation in US Toxic Substances Control Act new chemical risk assessments. *Environmental Science: Processes & Impacts*, *19*, 203–212. doi: 10.1039/C7EM00064B.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. doi: 10.1145/2939672.2939785.

Chen, Z., Zhang, P., Brown, K. G., Branch, J. L., Van Der Sloot, H. A., Meeussen, J. C. L., Delapp, R. C., Um, W., & Kosson, D. S. (2021). Development of a Geochemical Speciation Model for Use in Evaluating Leaching from a Cementitious Radioactive Waste Form. *Environmental Science & Technology*, *55*, 8642–8653. doi: 10.1021/acs.est.0c06227.

Chu, H., & Lu, W. (2015). Optimization design based on ensemble surrogate models for DNAPLs-contaminated groundwater remediation. *Journal of Water Supply: Research and Technology-Aqua*, *64*, 697–707. doi: 10.2166/aqua.2015.089.

Comans, R., & Roskam, G. (2002). Leaching procedure for the availability of polycyclic aromatic hydrocarbons (PAHs) in contaminated soil and waste materials. In P. Quevauviller (Ed.), *Methodologies in soil and sediment fractionation studies: Single and sequential extraction procedures* (pp. 123–141). Royal Society of Chemistry. doi: https://doi.org/10.1039/9781847551412.

Corwin, D. L. (2000). Evaluation of a simple lysimeter-design modification to minimize sidewall flow. *Journal of Contaminant Hydrology*, *42*, 35–49. doi: 10.1016/S0169-7722(99)00088-1.

Cousins, I. T., DeWitt, J. C., Glüge, J., Goldenman, G., Herzke, D., Lohmann, R., Miller, M., Ng, C. A., Scheringer, M., Vierke, L., & Wang, Z. (2020). Strategies for grouping per- and polyfluoroalkyl substances (PFAS) to protect human and environmental health. *Environmental Science: Processes & Impacts*, *22*, 1444–1460. doi: 10.1039/D0EM00147C.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*, 30055–30062. doi: 10.1073/pnas.1912789117.

Dąbrowska, D., Sołtysiak, M., Biniecka, P., Michalska, J., Wasilkowski, D., Nowak, A., & Nourani, V. (2019). Application of hydrogeological and biological research for the lysimeter experiment performance under simulated municipal landfill condition. *Journal of Material Cycles and Waste Management*, *21*, 1477–1487. doi: 10.1007/s10163-019-00900-x.

D'Agostino, R., & Pearson, E. S. (1973). Tests for Departure from Normality. Empirical Results for the Distributions of b 2 and b 1. *Biometrika*, *60*, 613. doi: 10.2307/2335012.

Dahlbo, H., et al. (2015). Construction and demolition waste management – a holistic overview. *Journal of Cleaner Production*, *107*, 610–618. doi: 10.1016/j.jclepro.2015.02.073.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*, 1–18. doi: 10.1016/S0169-7439(99)00047-7.

Deistler, M., Goncalves, P. J., & Macke, J. H. (2022). Truncated proposals for scalable and hassle-free simulation-based inference. doi: 10.48550/ARXIV.2210.04815.

Del Rey, I., Ayuso, J., Galvín, A., Jiménez, J., López, M., & García-Garrido, M. (2015). Analysis of chromium and sulphate origins in construction recycled materials based on leaching test results. *Waste Management*, *46*, 278–286. doi: 10.1016/j.wasman.2015.07.051.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In G. Goos, J. Hartmanis, & J. van Leeuwen (Eds.), *Multiple Classifier Systems* (pp. 1–15, Vol. 1857). Springer Berlin Heidelberg. doi: 10.1007/3-540-45014-9_1.

DIN 19528. (2009). *Elution von Feststoffen Perkolationsverfahren zur gemeinsamen Untersuchung des Elutionsverhaltens von anorganischen und organischen Stoffen; Ausgabe 01/2009; Beuth: Berlin, Germany* (tech. rep.). Beuth Verlag GmbH. doi: 10.31030/1399959.

DIN 19529. (2015). *Elution von Feststoffen_- Schüttelverfahren zur Untersuchung des Elutionsverhaltens von anorganischen und organischen Stoffen mit einem Wasser/Feststoff-Verhältnis von_2 l/kg; Beuth: Berlin, Germany* (tech. rep.). Beuth Verlag GmbH. doi: 10.31030/2359564.

DIN 38414. (1984). *Deutsche einheitsverfahren zur wasser-, abwasser- und schlammuntersuchung; schlamm und sedimente (gruppe s), bestimmung der eluierbarkeit mit wasser (dev s4)* (tech. rep.). Normenausschuss Wasserwesen, Beuth Verlag.

DIN 38414-6. (1986). *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung; Schlamm und Sedimente (Gruppe_s); Bestimmung der Sauerstoffverbrauchsrate (S_6); Beuth: Berlin, Germany* (tech. rep.). Beuth Verlag GmbH. doi: 10.31030/2007945.

DIN CEN/TS 16637-2. (2014). *Bauprodukte - Bewertung der Freisetzung von gefährlichen Stoffen_- Teil_2: Horizontale dynamische Oberflächenauslaugprüfung; Deutsche Fassung CEN/TS_16637-2:2014; Beuth: Berlin, Germany* (tech. rep.). Beuth Verlag GmbH. doi: 10.31030/2021527.

DIN EN 12457-1. (2003). *Charakterisierung von Abfällen_- Auslaugung; übereinstimmungsuntersuchung für die Auslaugung von körnigen Abfällen und Schlämmen_- Teil_1: Einstufiges Schüttelverfahren mit einem Flüssigkeits-/Feststoffverhältnis von 2_l/kg und einer Korngröße unter 4_mm (ohne oder mit Korngrößenreduzierung); Deutsche Fassung EN_12457-1:2002; Beuth: Berlin, Germany* (tech. rep.). Beuth Verlag GmbH. doi: 10.31030/9274065.

Dinno, A. (2015). Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. *The Stata Journal: Promoting communications on statistics and Stata*, *15*, 292–300. doi: 10.1177/1536867X1501500117.

Diotti, A., Perèz Galvin, A., Piccinali, A., Plizzari, G., & Sorlini, S. (2020). Chemical and Leaching Behavior of Construction and Demolition Wastes and Recycled Aggregates. *Sustainability*, *12*, 10326. doi: 10.3390/su122410326.

Dorj, E., & Altangerel, E. (2013). Anomaly detection approach using Hidden Markov Model. *Ifost*, 141–144. doi: 10.1109/IFOST.2013.6616874.

Doucette, W. J. (2003). Quantitative structure-activity relationships for predicting soil-sediment sorption coefficients for organic chemicals. *Environmental Toxicology and Chemistry*, *22*, 1771–1788. doi: 10.1897/01-362.

Du, Z., Deng, S., Bei, Y., Huang, Q., Wang, B., Huang, J., & Yu, G. (2014). Adsorption behavior and mechanism of perfluorinated compounds on various adsorbents—A review. *Journal of Hazardous Materials*, *274*, 443–454. doi: 10.1016/j.jhazmat.2014.04.038.

Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, *503*, 92–108. doi: 10.1016/j.neucom.2022.06.111.

Efron, B., & Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman; Hall/CRC. doi: 10.1201/9780429246593.

Engelsen, C. J., van der Sloot, H. A., & Petkovic, G. (2017). Long-term leaching from recycled concrete aggregates applied as sub-base material in road construction. *Science of The Total Environment*, *587-588*, 94–101. doi: 10.1016/j.scitotenv.2017.02.052.

Engelsen, C. J., Wibetoe, G., van der Sloot, H. A., Lund, W., & Petkovic, G. (2012). Field site leaching from recycled concrete aggregates applied as sub-base material in road construction. *Science of The Total Environment*, *427-428*, 86–97. doi: 10.1016/j.scitotenv.2012.04.021.

Ershadi, A., Finkel, M., Liu, B., Cirpka, O. A., & Grathwohl, P. (2024). Ensemble surrogate modeling of advective-dispersive transport with intraparticle diffusion model for column-leaching test. *Journal of Contaminant Hydrology*, *267*, 104423. doi: 10.1016/j.jconhyd.2024.104423.

Ershadi, A., Finkel, M., Susset, B., & Grathwohl, P. (2023). Applicability of machine learning models for the assessment of long-term pollutant leaching from solid waste materials. *Waste Management*, *171*, 337–349. doi: 10.1016/j.wasman.2023.09.001.

European Commission. (2014). *Energy, transport and environment indicators: 2014 edition.* Publications Office. Retrieved December 21, 2022, from https://data.europa.eu/doi/10.2785/56625

European Commission. (2018). *COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A European Strategy for Plastics in a Circular Economy*. Publications Office. Retrieved December 1, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:28:FIN

Fabregat-Palau, J., Rigol, A., Grathwohl, P., & Vidal, M. (2024). Assessing sorption of fluoroquinolone antibiotics in soils from a Kd compilation based on pure organic and mineral components. *Ecotoxicology and Environmental Safety*, *280*, 116535. doi: 10.1016/j.ecoenv.2024.116535.

Fabregat-Palau, J., Vidal, M., & Rigol, A. (2021). Modelling the sorption behaviour of perfluoroalkyl carboxylates and perfluoroalkane sulfonates in soils. *Science of The Total Environment*, *801*, 149343. doi: 10.1016/j.scitotenv.2021.149343.

Fabregat-Palau, J., Vidal, M., & Rigol, A. (2022). Examining sorption of perfluoroalkyl substances (PFAS) in biochars and other carbon-rich materials. *Chemosphere*, *302*, 134733. doi: 10.1016/j.chemosphere.2022.134733.

Fabregat-Palau, J., Yu, Z., Zeng, X., Vidal, M., & Rigol, A. (2023). Deriving parametric and probabilistic Kd values for fluoroquinolones in soils. *Science of The Total Environment*, *861*, 160266. doi: 10.1016/j.scitotenv.2022.160266.

Faucheux, L., Resche-Rigon, M., Curis, E., Soumelis, V., & Chevret, S. (2021). Clustering with missing and left-censored data: A simulation study comparing multiple-imputation-based procedures. *Biometrical Journal*, *63*, 372–393. doi: 10.1002/bimj.201900366.

Fenton, S. E., Ducatman, A., Boobis, A., DeWitt, J. C., Lau, C., Ng, C., Smith, J. S., & Roberts, S. M. (2021). Per- and Polyfluoroalkyl Substance Toxicity and Human Health Review: Current State of Knowledge and Strategies for Informing Future Research. *Environmental Toxicology and Chemistry*, *40*, 606–630. doi: 10.1002/etc.4890.

Feurer, M., & Hutter, F. (2019). Automated Machine Learning, Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.). Springer International Publishing. doi: 10.1007/978-3-030-05318-5_1.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and Robust Automated Machine Learning (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett, Eds.). *Advances in Neural Information Processing Systems*, *28*. https://proceedings.neurips.cc/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf

Finkel, M., & Grathwohl, P. (2017). Impact of pre-equilibration and diffusion limited release kinetics on effluent concentration in column leaching tests: Insights from numerical simulations. *Waste Management*, *63*, 58–73. doi: 10.1016/j.wasman.2016.11.031.

Flores, V., Keith, B., & Leiva, C. (2020). Using Artificial Intelligence Techniques to Improve the Prediction of Copper Recovery by Leaching. *Journal of Sensors*, *2020*, e2454875. doi: 10.1155/2020/2454875.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*. doi: 10.1214/aos/1013203451.

Games, P. A., & Howell, J. F. (1976). Pairwise Multiple Comparison Procedures with Unequal N's and/or Variances: A Monte Carlo Study. *Journal of Educational Statistics*, *1*, 113–125. doi: 10.3102/10769986001002113.

Garud, S. S., Karimi, I., & Kraft, M. (2017). Smart Sampling Algorithm for Surrogate Model Development. *Computers & Chemical Engineering*, *96*, 103–114. doi: 10.1016/j.compchemeng.2016.10.006.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3–42. doi: 10.1007/s10994-006-6226-1.

Ghahramani, Z. (2004). Unsupervised Learning. In O. Bousquet, U. Von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning* (pp. 72–112, Vol. 3176). Springer Berlin Heidelberg. doi: 10.1007/978-3-540-28650-9_5.

Glüge, J., Scheringer, M., Cousins, I. T., DeWitt, J. C., Goldenman, G., Herzke, D., Lohmann, R., Ng, C. A., Trier, X., & Wang, Z. (2020). An overview of the uses of per- and polyfluoroalkyl substances (PFAS). *Environmental Science: Processes & Impacts*, *22*, 2345–2373. doi: 10.1039/D0EM00291G.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT press.

Grathwohl, P., & Susset, B. (2009). Comparison of percolation to batch and sequential leaching tests: Theory and data. *Waste Management*, *29*, 2681–2688. doi: 10.1016/j.wasman.2009.05.016.

Grathwohl, P. (1990). Influence of organic matter from soils and sediments from various origins on the sorption of some chlorinated aliphatic hydrocarbons: Implications on Koc correlations. *Environmental Science & Technology*, *24*, 1687–1693. doi: 10.1021/es00081a010.

Grathwohl, P. (2014). On equilibration of pore water in column leaching tests. *Waste Management*, *34*, 908–918. doi: 10.1016/j.wasman.2014.02.012.

Grathwohl, P., & Sloot, H. v. d. (2007). Groundwater Risk Assessment at Contaminated Sites (GRACOS): Test Methods and Modelling Approaches. In P. Quevauviller (Ed.), *Groundwater Science and Policy* (pp. 291–315). Royal Society of Chemistry. doi: 10.1039/9781847558039-00291.

Greenberg, D., Nonnenmacher, M., & Macke, J. (2019). Automatic Posterior Transformation for Likelihood-Free Inference. *Proceedings of the 36th International Conference on Machine Learning*, 2404–2414. Retrieved January 30, 2023, from https://proceedings.mlr.press/v97/greenberg19a.html

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A Simple and Effective Model-Based Variable Importance Measure. doi: 10.48550/ARXIV.1805.04755.

Guelfo, J. L., & Higgins, C. P. (2013). Subsurface Transport Potential of Perfluoroalkyl Acids at Aqueous Film-Forming Foam (AFFF)-Impacted Sites. *Environmental Science & Technology*, *47*, 4164–4171. doi: 10.1021/es3048043.

Guo, B., Zeng, J., & Brusseau, M. L. (2020). A Mathematical Model for the Release, Transport, and Retention of Per- and Polyfluoroalkyl Substances (PFAS) in the Vadose Zone. *Water Resources Research*, *56*, e2019WR026667. doi: 10.1029/2019WR026667.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, *2*, 84–90. doi: 10.1007/BF01386213.

Hastie, T. (2020). Ridge Regularization: An Essential Concept in Data Science. *Technometrics*, *62*, 426–433. doi: 10.1080/00401706.2020.1791959.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

Hazra, T., & Anjaria, K. (2022). Applications of game theory in deep learning: A survey. *Multimedia Tools and Applications*, *81*, 8963–8994. doi: 10.1007/s11042-022-12153-2.

Hbali, Y., Hbali, S., Ballihi, L., & Sadgal, M. (2018). Skeleton-based human activity recognition for elderly monitoring systems. *IET Computer Vision*, *12*, 16–26. doi: 10.1049/iet-cvi.2017.0062.

Higgins, C. P., & Luthy, R. G. (2006). Sorption of Perfluorinated Surfactants on Sediments. *Environmental Science & Technology*, *40*, 7251–7256. doi: 10.1021/es061000n.

Higgins, C. P., & Luthy, R. G. (2007). Modeling Sorption of Anionic Surfactants onto Sediment Materials: An a priori Approach for Perfluoroalkyl Surfactants and Linear Alkylbenzene Sulfonates. *Environmental Science & Technology*, *41*, 3254–3261. doi: 10.1021/es062449j.

Hodson, J., & Williams, N. (1988). The estimation of the adsorption coefficient (Koc) for soils by high performance liquid chromatography. *Chemosphere*, *17*, 67–77. doi: 10.1016/0045-6535(88)90045-8.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*, 55–67. doi: 10.1080/00401706.1970.10488634.

Huang, B., Li, Z., Huang, J., Chen, G., Nie, X., Ma, W., Yao, H., Zhen, J., & Zeng, G. (2015). Aging effect on the leaching behavior of heavy metals (Cu, Zn, and Cd) in red paddy soil. *Environmental Science and Pollution Research*, *22*, 11467–11477. doi: 10.1007/s11356-015-4386-x.

Huang, Y., Lu, M., Li, H., Bai, M., & Huang, X. (2019). Sensitive determination of perfluoroalkane sulfonamides in water and urine samples by multiple monolithic fiber solid-phase microextraction and liquid chromatography tandem mass spectrometry. *Talanta*, *192*, 24–31. doi: 10.1016/j.talanta.2018.09.004.

Hyks, J., Astrup, T., & Christensen, T. H. (2009). Leaching from MSWI bottom ash: Evaluation of non-equilibrium in column percolation experiments. *Waste Management*, *29*, 522–529. doi: 10.1016/j.wasman.2008.06.011.

Imoto, Y. (2024). Insight into the relationship between similarity and the degree of equilibrium of contaminant release curves through numerical simulations. *Journal of Contaminant Hydrology*, *267*, 104451. doi: 10.1016/j.jconhyd.2024.104451.

Inoue, Y., Hashizume, N., Yakata, N., Murakami, H., Suzuki, Y., Kikushima, E., & Otsuka, M. (2012). Unique Physicochemical Properties of Perfluorinated Compounds and Their Bioconcentration in Common Carp Cyprinus carpio L. *Archives of Environmental Contamination and Toxicology*, *62*, 672–680. doi: 10.1007/s00244-011-9730-7.

Jaeger, R., & Liedl, R. (2000). Prognose der Sorptionskinetik organischer Schadstoffe in heterogenem Aquifermaterial. *Grundwasser*, *5*, 57–66. doi: 10.1007/s767-000-8348-2.

Jang, Y.-C., & Townsend, T. (2001). Sulfate leaching from recovered construction and demolition debris fines. *Advances in Environmental Research*, *5*, 203–217. doi: 10.1016/S1093-0191(00)00056-3.

Jha, D., Choudhary, K., Tavazza, F., Liao, W.-k., Choudhary, A., Campbell, C., & Agrawal, A. (2019). Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature Communications*, *10*, 5316. doi: 10.1038/s41467-019-13297-w.

Jiang, X., Lu, W., Hou, Z., Zhao, H., & Na, J. (2015). Ensemble of surrogates-based optimization for identifying an optimal surfactant-enhanced aquifer remediation strategy at heterogeneous DNAPL-contaminated sites. *Computers & Geosciences*, *84*, 37–45. doi: 10.1016/j.cageo.2015.08.003.

Johnsen, A. R., Wick, L. Y., & Harms, H. (2005). Principles of microbial PAH-degradation in soil. *Environmental Pollution*, *133*, 71–84. doi: 10.1016/j.envpol.2004.04.015.

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*, 455–492. doi: 10.1023/A:1008306431147.

Kalbe, U., Bandow, N., Bredow, A., Mathies, H., & Piechotta, C. (2014). Column leaching tests on soils containing less investigated organic pollutants. *Journal of Geochemical Exploration*, *147*, 291–297. doi: 10.1016/j.gexplo.2014.06.014.

Kalbe, U., Berger, W., Eckardt, J., & Simon, F.-G. (2008). Evaluation of leaching and extraction procedures for soil and waste. *Waste Management*, *28*, 1027–1038. doi: 10.1016/j.wasman.2007.03.008.

Kalbe, U., Berger, W., Simon, F.-G., Eckardt, J., & Christoph, G. (2007). Results of interlaboratory comparisons of column percolation tests. *Journal of Hazardous Materials*, *148*, 714–720. doi: 10.1016/j.jhazmat.2007.03.039.

Kenaga, E., & Goring, C. (1980). Relationship Between Water Solubility, Soil Sorption, Octanol-Water Partitioning, and Concentration of Chemicals in Biota. In *Aquatic Toxicology* (pp. 78–115). ASTM International100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959. doi: 10.1520/STP27410S.

Khan, A., Hassett, J. J., & Banwart, W. L. (1979). SORPTION OF ACETOPHENONE BY SEDIMENTS AND SOILS: *Soil Science*, *128*, 297–302. doi: 10.1097/00010694-197911000-00007.

Kleefeld, A., & Reißel, M. (2011). The Levenberg–Marquardt method applied to a parameter estimation problem arising from electrical resistivity tomography. *Applied Mathematics and Computation*, *217*, 4490–4501. doi: 10.1016/j.amc.2010.10.052.

Kleineidam, S., Schüth, C., & Grathwohl, P. (2002). Solubility-Normalized Combined Adsorption-Partitioning Sorption Isotherms for Organic Pollutants. *Environmental Science & Technology*, *36*, 4689–4697. doi: 10.1021/es010293b.

Knight, E. R., Janik, L. J., Navarro, D. A., Kookana, R. S., & McLaughlin, M. J. (2019). Predicting partitioning of radiolabelled 14C-PFOA in a range of soils using diffuse reflectance infrared spectroscopy. *Science of The Total Environment*, *686*, 505–513. doi: 10.1016/j.scitotenv.2019.05.339.

Kobyzev, I., Prince, S. J., & Brubaker, M. A. (2021). Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*, 3964–3979. doi: 10.1109/TPAMI.2020.2992934.

Kreislaufwirtschaft Bau. (2018). *Mineralische Bauabfälle Monitoring. Bericht zum Aufkommen und zum Verbleib mineralischer Bauabfälle* (tech. rep.). Kreislaufwirtschaft BAU: Berlin, Germany. Retrieved December 19, 2022, from https://www.kreislaufwirtschaft-bau.de/Download/Bericht-12.pdf

Langberg, H. A., Arp, H. P. H., Breedveld, G. D., Slinde, G. A., Høiseter, A., Grønning, H. M., Jartun, M., Rundberget, T., Jenssen, B. M., & Hale, S. E. (2021). Paper product production identified as the main source of per- and polyfluoroalkyl substances (PFAS) in a Norwegian lake: Source and historic emission tracking. *Environmental Pollution*, *273*, 116259. doi: 10.1016/j.envpol.2020.116259.

Lee, H., & Mabury, S. A. (2011). A Pilot Survey of Legacy and Current Commercial Fluorinated Chemicals in Human Sera from United States Donors in 2009. *Environmental Science & Technology*, *45*, 8067–8074. doi: 10.1021/es200167q.

Lee, H., & Mabury, S. A. (2017). Sorption of Perfluoroalkyl Phosphonates and Perfluoroalkyl Phosphinates in Soils. *Environmental Science & Technology*, *51*, 3197–3205. doi: 10.1021/acs.est.6b04395.

Lepot, M., Aubin, J.-B., & Clemens, F. (2017). Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment. *Water*, *9*, 796. doi: 10.3390/w9100796.

Lever, J., Krzywinski, M., & Altman, N. (2016). Model selection and overfitting. *Nature Methods*, *13*, 703–704. doi: 10.1038/nmeth.3968.

Li, F., Fang, X., Zhou, Z., Liao, X., Zou, J., Yuan, B., & Sun, W. (2019). Adsorption of perfluorinated acids onto soils: Kinetics, isotherms, and influences of soil properties. *Science of The Total Environment*, *649*, 504–514. doi: 10.1016/j.scitotenv.2018.08.209.

Li, L., Kumar Damarla, S., Wang, Y., & Huang, B. (2021). A Gaussian mixture model based virtual sample generation approach for small datasets in industrial processes. *Information Sciences*, *581*, 262–277. doi: 10.1016/j.ins.2021.09.014.

Li, Y., Oliver, D. P., & Kookana, R. S. (2018). A critical analysis of published data to discern the role of soil and sediment properties in determining sorption of per and polyfluoroalkyl substances (PFASs). *Science of The Total Environment*, *628-629*, 110–120. doi: 10.1016/j.scitotenv.2018.01.167.

Liang, J., Li, Z., Pan, L., Khailah, E. Y., Sun, L., & Lu, W. (2023). Research on surrogate model of dam numerical simulation with multiple outputs based on adaptive sampling. *Scientific Reports*, *13*, 11955. doi: 10.1038/s41598-023-38590-z.

Lillington, J. N., Goût, T. L., Harrison, M. T., & Farnan, I. (2020). Assessing static glass leaching predictions from large datasets using machine learning. *Journal of Non-Crystalline Solids*, *546*, 120276. doi: 10.1016/j.jnoncrysol.2020.120276.

Lin, X., Vollpracht, A., Markus, P., & Linnemann, V. (2020). Optimization of a German short term percolation test to determine the leaching of granular materials. *Waste Management*, *105*, 433–444. doi: 10.1016/j.wasman.2020.02.039.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*, 319–330. doi: 10.1002/asmb.446.

Liu, B., Finkel, M., & Grathwohl, P. (2021). Mass Transfer Principles in Column Percolation Tests: Initial Conditions and Tailing in Heterogeneous Materials. *Materials*, *14*, 4708. doi: 10.3390/ma14164708.

Liu, B., Finkel, M., & Grathwohl, P. (2022a). First order approximation for coupled film and intraparticle pore diffusion to model sorption/desorption batch experiments. *Journal of Hazardous Materials*, *429*, 128314. doi: 10.1016/j.jhazmat.2022.128314.

Liu, M., Munoz, G., Vo Duy, S., Sauvé, S., & Liu, J. (2022b). Per- and Polyfluoroalkyl Substances in Contaminated Soil and Groundwater at Airports: A Canadian Case Study. *Environmental Science & Technology*, *56*, 885–895. doi: 10.1021/acs.est.1c04798.

Løkke, H. (1984). Sorption of selected organic pollutants in Danish soils. *Ecotoxicology and Environmental Safety*, *8*, 395–409. doi: 10.1016/0147-6513(84)90062-9.

López-Uceda, A., Galvín, A. P., Barbudo, A., & Ayuso, J. (2019). Long-term leaching and mechanical behaviour at recycled aggregate with different gypsum contents. *Environmental Science and Pollution Research*, *26*, 35565–35573. doi: 10.1007/s11356-019-04925-5.

Löv, A., Larsbo, M., Sjöstedt, C., Cornelis, G., Gustafsson, J. P., & Kleja, D. B. (2019). Evaluating the ability of standardised leaching tests to predict metal(loid) leaching from intact soil columns using size-based elemental fractionation. *Chemosphere*, *222*, 453–460. doi: 10.1016/j.chemosphere.2019.01.148.

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., & Macke, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. *Advances in Neural Information Processing Systems*, *30*. doi: https://doi.org/10.48550/arXiv.1711.01861.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Computing Research Repository*, *abs/1705.07874*. doi: 10.48550/ARXIV.1705.07874.

Lupsea, M., Tiruta-Barna, L., & Schiopu, N. (2014). Leaching of hazardous substances from a composite construction product - An experimental and modelling approach for fibre-cement sheets. *Journal of Hazardous Materials*, *264*, 236–245. doi: 10.1016/j.jhazmat.2013.11.017.

Ma, J., & Yuan, Y. (2019). Dimension reduction of image deep feature using PCA. *Journal of Visual Communication and Image Representation*, *63*, 102578. doi: 10.1016/j.jvcir.2019.102578.

Ma, W., Wang, M., Jiang, R., & Chen, W. (2023). A machine learning based approach for estimating site-specific partition coefficient Kd of organic compounds: Application to nonionic pesticides. *Environmental Pollution*, *323*, 121297. doi: 10.1016/j.envpol.2023.121297.

Mahalanobis, P. C. (2018). ON THE GENERALIZED DISTANCE IN STATISTICS. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, *80*, S1–S7. Retrieved January 4, 2024, from https://www.jstor.org/stable/48723335

McKight, P. E., & Najab, J. (2010). Kruskal-Wallis Test. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini Encyclopedia of Psychology* (1st ed., pp. 1–1). Wiley. doi: 10.1002/9780470479216.corpsy0491.

Meima, J. A., & Comans, R. N. (1999). The leaching of trace elements from municipal solid waste incinerator bottom ash at different stages of weathering. *Applied Geochemistry*, *14*, 159–171. doi: 10.1016/S0883-2927(98)00047-X.

Mejia-Avendaño, S., Zhi, Y., Yan, B., & Liu, J. (2020). Sorption of Polyfluoroalkyl Surfactants on Surface Soils: Effect of Molecular Structures, Soil Properties, and Solution Chemistry. *Environmental Science & Technology*, *54*, 1513–1521. doi: 10.1021/acs.est.9b04989.

Mertens, J., & Smolders, E. (2013). Zinc. In B. J. Alloway (Ed.), *Heavy Metals in Soils* (pp. 465–493, Vol. 22). Springer Netherlands. doi: 10.1007/978-94-007-4470-7_17.

Mesquita, M. E., & Carranca, C. (2005). Effect of Dissolved Organic Matter on Copper - Zinc Competitive Adsorption by a Sandy Soil at Different pH Values. *Environmental Technology*, *26*, 1065–1072. doi: 10.1080/09593332608618493.

Milinovic, J., Lacorte, S., Vidal, M., & Rigol, A. (2015). Sorption behaviour of perfluoroalkyl substances in soils. *Science of The Total Environment*, *511*, 63–71. doi: 10.1016/j.scitotenv.2014.12.017.

Moghadasi, R., Mumberg, T., & Wanner, P. (2023). Spatial Prediction of Concentrations of Per- and Polyfluoroalkyl Substances (PFAS) in European Soils. *Environmental Science & Technology Letters*, *10*, 1125–1129. doi: 10.1021/acs.estlett.3c00633.

Mohammadi, F., Eggenweiler, E., Flemisch, B., Oladyshkin, S., Rybak, I., Schneider, M., & Weishaupt, K. (2023). A surrogate-assisted uncertainty-aware Bayesian validation framework and its application to coupling free flow and porous-medium flow. *Computational Geosciences*, *27*, 663–686. doi: 10.1007/s10596-023-10228-z.

Molenaar, A. A. A., & van Niekerk, A. A. (2002). Effects of Gradation, Composition, and Degree of Compaction on the Mechanical Characteristics of Recycled Unbound Materials. *Transportation Research Record: Journal of the Transportation Research Board*, *1787*, 73–82. doi: 10.3141/1787-08.

Moré, J. J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. In G. A. Watson (Ed.), *Numerical Analysis* (pp. 105–116, Vol. 630). Springer Berlin Heidelberg. doi: 10.1007/BFb0067700.

Moss, G., Višnjević, V., Eisen, O., Oraschewski, F. M., Schröder, C., Macke, J. H., & Drews, R. (2023). Simulation-Based Inference of Surface Accumulation and Basal Melt Rates of an Antarctic Ice Shelf from Isochronal Layers. doi: 10.48550/ARXIV.2312.02997.

Munoz, G., Liu, J., Vo Duy, S., & Sauvé, S. (2019). Analysis of F-53B, Gen-X, ADONA, and emerging fluoroalkylether substances in environmental and biomonitoring samples: A review. *Trends in Environmental Analytical Chemistry*, *23*, e00066. doi: 10.1016/j.teac.2019.e00066.

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: An overview, II. *WIREs Data Mining and Knowledge Discovery*, *7*, e1219. doi: 10.1002/widm.1219.

Naka, A., Yasutaka, T., Sakanakura, H., Kalbe, U., Watanabe, Y., Inoba, S., Takeo, M., Inui, T., Katsumi, T., Fujikawa, T., Sato, K., Higashino, K., & Someya, M. (2016). Column percolation test for contaminated soils: Key factors for standardization. *Journal of Hazardous Materials*, *320*, 326–340. doi: 10.1016/j.jhazmat.2016.08.046.

NEN 7349. (1995). *Leaching characteristics of solid earthy and stony building and waste materials. Leaching tests: Determination of the leaching behaviour of inorganic components from granular waste with the cascade test; Netherlands Normalisatie Institut, Delft* (tech. rep.). Nederlands Normalisatie institut, Delft. https://www.nen.nl/en/norm/pdf/preview/document/12942/

Nguyen, T. M. H., Bräunig, J., Thompson, K., Thompson, J., Kabiri, S., Navarro, D. A., Kookana, R. S., Grimison, C., Barnes, C. M., Higgins, C. P., McLaughlin, M. J., & Mueller, J. F. (2020). Influences of Chemical Properties, Soil Properties, and Solution pH on Soil–Water Partitioning Coefficients of Per- and Polyfluoroalkyl Substances (PFASs). *Environmental Science & Technology*, *54*, 15883–15892. doi: 10.1021/acs.est.0c05705.

OECD. (2000). *Test No. 106: Adsorption – Desorption Using a Batch Equilibrium Method*. doi: 10.1787/9789264069602-en.

Ogunsanya, M., Isichei, J., & Desai, S. (2023). Grid search hyperparameter tuning in additive manufacturing processes. *Manufacturing Letters*, *35*, 1031–1042. doi: 10.1016/j.mfglet.2023.08.056.

Oladyshkin, S., & Nowak, W. (2019). The Connection between Bayesian Inference and Information Theory for Model Selection, Information Gain and Experimental Design. *Entropy*, *21*, 1081. doi: 10.3390/e21111081.

Oliver, D. P., Li, Y., Orr, R., Nelson, P., Barnes, M., McLaughlin, M. J., & Kookana, R. S. (2019). The role of surface charge and pH changes in tropical soils on sorption behaviour of per- and polyfluoroalkyl substances (PFASs). *Science of The Total Environment*, *673*, 197–206. doi: 10.1016/j.scitotenv.2019.04.055.

Oorts, K. (2013). Copper. In B. J. Alloway (Ed.), *Heavy Metals in Soils* (pp. 367–394, Vol. 22). Springer Netherlands. doi: 10.1007/978-94-007-4470-7_13.

Ouyang, Q., Lu, W., Lin, J., Deng, W., & Cheng, W. (2017). Conservative strategy-based ensemble surrogate model for optimal groundwater remediation design at DNAPLs-contaminated sites. *Journal of Contaminant Hydrology*, *203*, 1–8. doi: 10.1016/j.jconhyd.2017.05.007.

Papamakarios, G., & Murray, I. (2016). Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation. *Advances in Neural Information Processing Systems*, *29*. doi: https://doi.org/10.48550/arXiv.1605.06376.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2019). Normalizing Flows for Probabilistic Modeling and Inference. doi: 10.48550/ARXIV.1912.02762.

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. doi: 10.48550/ARXIV.1503.06462.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2012). Scikit-learn: Machine Learning in Python. *Computing Research Repository*. doi: 10.48550/ARXIV.1201.0490.

Peng, H. (2019). A literature review on leaching and recovery of vanadium. *Journal of Environmental Chemical Engineering*, *7*, 103313. doi: 10.1016/j.jece.2019.103313.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, *7*, 217–240. doi: 10.5194/soil-7-217-2021.

Prado, F. E., Hilal, M., Chocobar-Ponce, S., Pagano, E., Rosa, M., & Prado, C. (2016). Chromium and the Plant. In *Plant Metal Interaction* (pp. 149–177). Elsevier. doi: 10.1016/B978-0-12-803158-2.00006-0.

Prasad, A., Suggala, A. S., Balakrishnan, S., & Ravikumar, P. (2020). Robust Estimation via Robust Gradient Estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*, 601–627. doi: 10.1111/rssb.12364.

Prevedouros, K., Cousins, I. T., Buck, R. C., & Korzeniowski, S. H. (2006). Sources, Fate and Transport of Perfluorocarboxylates. *Environmental Science & Technology*, *40*, 32–44. doi: 10.1021/es0512475.

Prieto-Espinoza, M., Susset, B., & Grathwohl, P. (2022). Long-Term Leaching Behavior of Organic and Inorganic Pollutants after Wet Processing of Solid Waste Materials. *Materials*, *15*, 858. doi: 10.3390/ma15030858.

Rahman, M. A., Imteaz, M., Arulrajah, A., & Disfani, M. M. (2014). Suitability of recycled construction and demolition aggregates as alternative pipe backfilling materials. *Journal of Cleaner Production*, *66*, 75–84. doi: 10.1016/j.jclepro.2013.11.005.

Rahman, M. A., Imteaz, M. A., Arulrajah, A., Piratheepan, J., & Disfani, M. M. (2015). Recycled construction and demolition materials in permeable pavement systems: Geotechnical and hydraulic characteristics. *Journal of Cleaner Production*, *90*, 183–194. doi: 10.1016/j.jclepro.2014.11.042.

Rankin, K., Mabury, S. A., Jenkins, T. M., & Washington, J. W. (2016). A North American and global survey of perfluoroalkyl substances in surface soils: Distribution patterns and mode of occurrence. *Chemosphere*, *161*, 333–341. doi: 10.1016/j.chemosphere.2016.06.109.

Rayne, S., & Forest, K. (2009). Comment on "Indirect Photolysis of Perfluorochemicals: Hydroxyl Radical-Initiated Oxidation of *N* -Ethyl Perfluorooctane Sulfonamido Acetate (*N*-EtFOSAA) and Other Perfluoroalkanesulfonamides". *Environmental Science & Technology*, *43*, 7995–7996. doi: 10.1021/es9022464.

Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, *48*. doi: 10.1029/2011WR011527.

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-Validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 1–7). Springer New York. doi: 10.1007/978-1-4899-7993-3_565-2.

Regis, R. G., & Shoemaker, C. A. (2007). A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions. *INFORMS Journal on Computing*, *19*, 497–509. doi: 10.1287/ijoc.1060.0182.

Regis, R. G., & Shoemaker, C. A. (2009). Parallel Stochastic Global Optimization Using Radial Basis Functions. *INFORMS Journal on Computing*, *21*, 411–426. doi: 10.1287/ijoc.1090.0325.

Rehder, D. (2015). The role of vanadium in biology. *Metallomics*, *7*, 730–742. doi: 10.1039/C4MT00304G.

Richey, D. G., Driscoll, C. T., & Likens, G. E. (1997). Soil Retention of Trifluoroacetate. *Environmental Science & Technology*, *31*, 1723–1727. doi: 10.1021/es960649x.

Rodrigues, F., Carvalho, M. T., Evangelista, L., & De Brito, J. (2013). Physical–chemical and mineralogical characterization of fine aggregates from construction and demolition waste recycling plants. *Journal of Cleaner Production*, *52*, 438–445. doi: 10.1016/j.jclepro.2013.02.023.

Röhler, K., Haluska, A. A., Susset, B., Liu, B., & Grathwohl, P. (2021). Long-term behavior of PFAS in contaminated agricultural soils in Germany. *Journal of Contaminant Hydrology*, *241*, 103812. doi: 10.1016/j.jconhyd.2021.103812.

Roussat, N., Mehu, J., Abdelghafour, M., & Brula, P. (2008). Leaching behaviour of hazardous demolition waste. *Waste Management*, *28*, 2032–2040. doi: 10.1016/j.wasman.2007.10.019.

Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.

Saeidi, N., Lai, A., Harnisch, F., & Sigmund, G. (2024). A FAIR comparison of activated carbon, biochar, cyclodextrins, polymers, resins, and metal organic frameworks for the adsorption of per- and polyfluorinated substances. *Chemical Engineering Journal*, *498*, 155456. doi: 10.1016/j.cej.2024.155456.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*, 160. doi: 10.1007/s42979-021-00592-x.

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning*, *10*, 153–178. doi: 10.1007/BF00993504.

Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and algorithms*. MIT Press. doi: https://doi.org/10.7551/mitpress/8291.001.0001.

Schellenberger, S., Jönsson, C., Mellin, P., Levenstam, O. A., Liagkouridis, I., Ribbenstedt, A., Hanning, A.-C., Schultes, L., Plassmann, M. M., Persson, C., Cousins, I. T., & Benskin, J. P. (2019). Release of Side-Chain Fluorinated Polymer-Containing Microplastic Fibers from Functional Textiles During Washing and First Estimates of Perfluoroalkyl Acid Emissions. *Environmental Science & Technology*, *53*, 14329–14338. doi: 10.1021/acs.est.9b04165.

Schlummer, M., Gruber, L., Fiedler, D., Kizlauskas, M., & Müller, J. (2013). Detection of fluorotelomer alcohols in indoor environments and their relevance for human exposure. *Environment International*, *57-58*, 42–49. doi: 10.1016/j.envint.2013.03.010.

Schüßler, M., Capitain, C., Bugsel, B., Zweigle, J., & Zwiener, C. (2024). Non-target screening reveals 124 PFAS at an AFFF-impacted field site in Germany specified by novel systematic terminology. *Analytical and Bioanalytical Chemistry*. doi: 10.1007/s00216-024-05611-3.

Shah, A. V., Srivastava, V. K., Mohanty, S. S., & Varjani, S. (2021). Municipal solid waste as a sustainable resource for energy production: State-of-the-art review. *Journal of Environmental Chemical Engineering*, *9*, 105717. doi: 10.1016/j.jece.2021.105717.

Shahid, M., Shamshad, S., Rafiq, M., Khalid, S., Bibi, I., Niazi, N. K., Dumat, C., & Rashid, M. I. (2017). Chromium speciation, bioavailability, uptake, toxicity and detoxification in soil-plant system: A review. *Chemosphere*, *178*, 513–533. doi: 10.1016/j.chemosphere.2017.03.074.

Shams, R., Alimohammadi, S., & Yazdi, J. (2021). Optimized stacking, a new method for constructing ensemble surrogate models applied to DNAPL-contaminated aquifer remediation. *Journal of Contaminant Hydrology*, *243*, 103914. doi: 10.1016/j.jconhyd.2021.103914.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611. doi: 10.1093/biomet/52.3-4.591.

Shapley, L. S. (1953). 17. A Value for n-Person Games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307–318). Princeton University Press. doi: 10.1515/9781400881970-018.

Sharma, P., Singh, S. P., Parakh, S. K., & Tong, Y. W. (2022). Health hazards of hexavalent chromium (Cr (VI)) and its microbial reduction. *Bioengineered*, *13*, 4923–4938. doi: 10.1080/21655979.2022.2037273.

Sheikholeslami, R., & Razavi, S. (2017). Progressive Latin Hypercube Sampling: An efficient approach for robust sampling-based analysis of environmental models. *Environmental Modelling & Software*, *93*, 109–126. doi: 10.1016/j.envsoft.2017.03.010.

Sheppard, S. C., Long, J. M., Sanipelli, B., & Sohlenius, G. (2009). Solid/liquid partition coefficients (kd) for selected soils and sediments at forsmark and laxemar-simpevarp. https://api.semanticscholar.org/CorpusID:127426902

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. doi: 10.1109/ACCESS.2020.2988796.

Smolders, E., & Mertens, J. (2013). Cadmium. In B. J. Alloway (Ed.), *Heavy Metals in Soils* (pp. 283–311, Vol. 22). Springer Netherlands. doi: 10.1007/978-94-007-4470-7_10.

Sobol', I. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, *7*, 86–112. doi: 10.1016/0041-5553(67)90144-9.

Spijker, J., Fraters, D., & Vrijhoef, A. (2021). A machine learning based modelling framework to predict nitrate leaching from agricultural soils across the Netherlands. *Environmental Research Communications*, *3*, 045002. doi: 10.1088/2515-7620/abf15f.

Su, G., Peng, L., & Hu, L. (2017). A Gaussian process-based dynamic surrogate model for complex engineering structural reliability analysis. *Structural Safety*, *68*, 97–109. doi: 10.1016/j.strusafe.2017.06.003.

Sundararajan, M., & Najmi, A. (2019). The many Shapley values for model explanation. doi: 10.48550/ARXIV.1908.08474.

Susset, B., & Grathwohl, P. (2011). Leaching standards for mineral recycling materials - A harmonized regulatory concept for the upcoming German Recycling Decree. *Waste Management*, *31*, 201–214. doi: 10.1016/j.wasman.2010.08.017.

Susset, B., & Leuchs, W. (2008a). *Umsetzung der Ergebnisse des BMBF-Verbundes Sickerwasser prognose in konkrete Vorschläge zur Harmonisierung von Methoden - Ableitung von Materialwerten im Eluat und Einbaumöglichkeiten mineralischer Ersatzbaustoffe. Fachbericht zum UBA-UFOPLAN-Vorhaben* (tech. rep.). UBA-UFOPLAN. https://www.umweltbundesamt.de/sites/default/files/medien/376/publikationen/4065.pdf

Tan, K., & Dowling, P. (1984). Effect of organic matter on CEC due to permanent and variable charges in selected temperate region soils. *Geoderma*, *32*, 89–101. doi: 10.1016/0016-7061(84)90065-X.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*, 363–377. doi: 10.1002/sam.11348.

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P., Greenberg, D., & Macke, J. (2020). Sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, *5*, 2505. doi: 10.21105/joss.02505.

Thang, T. M., & Kim, J. (2011). The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters. *2011 International Conference on Information Science and Applications*, 1–5. doi: 10.1109/ICISA.2011.5772437.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*, 520–525. doi: 10.1093/bioinformatics/17.6.520.

Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, *5*, 99. doi: 10.2307/3001913.

Umeh, A. C., Naidu, R., Shilpi, S., Boateng, E. B., Rahman, A., Cousins, I. T., Chadalavada, S., Lamb, D., & Bowman, M. (2021). Sorption of PFOS in 114 Well-Characterized Tropical and Temperate Soils: Application of Multivariate and Artificial Neural Network Analyses. *Environmental Science & Technology*, *55*, 1779–1789. doi: 10.1021/acs.est.0c07202.

US EPA Method 1314. (2013). *Liquid-solid partitioning as a function of liquid-to-solid ratio for constituents in solid materials using an up-flow percolation column procedure;United States Environmental Protection Agency, USA* (tech. rep.). Environmental Protection Agency. Retrieved December 19, 2022, from https://www.epa.gov/sites/default/files/2017-10/documents/method_1314_-_final_8-3-17.pdf

Van De Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*, 1. doi: 10.1038/s43586-020-00001-2.

van der Sloot, H. A. (2004). Readily accessible data and an integrated approach is needed for evaluating waste treatment options and preparation of materials for beneficial use. *Waste Management*, *24*, 751–752. doi: 10.1016/j.wasman.2004.08.001.

van Zomeren, A., & Comans, R. N. J. (2004). Contribution of Natural Organic Matter to Copper Leaching from Municipal Solid Waste Incinerator Bottom Ash. *Environmental Science & Technology*, *38*, 3927–3932. doi: 10.1021/es035266v.

Von Toussaint, U. (2011). Bayesian inference in physics. *Reviews of Modern Physics*, *83*, 943–999. doi: 10.1103/RevModPhys.83.943.

Vrugt, J. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling and Software*, *75*, 273–316. doi: 10.1016/j.envsoft.2015.08.013.

Wang, G. G., & Shan, S. (2007). Review of Metamodeling Techniques in Support of Engineering Design Optimization. *Journal of Mechanical Design*, *129*, 370–380. doi: 10.1115/1.2429697.

Wang, Z., Buser, A. M., Cousins, I. T., Demattio, S., Drost, W., Johansson, O., Ohno, K., Patlewicz, G., Richard, A. M., Walker, G. W., White, G. S., & Leinala, E. (2021). A New OECD Definition for Per- and Polyfluoroalkyl Substances. *Environmental Science & Technology*, *55*, 15575–15578. doi: 10.1021/acs.est.1c06896.

Wang, Z., MacLeod, M., Cousins, I. T., Scheringer, M., & Hungerbühler, K. (2011). Using COSMOtherm to predict physicochemical properties of poly- and perfluorinated alkyl substances (PFASs). *Environmental Chemistry*, *8*, 389. doi: 10.1071/EN10143.

Wei, C., Song, X., Wang, Q., Liu, Y., & Lin, N. (2019). Influence of coexisting Cr(VI) and sulfate anions and Cu(II) on the sorption of F-53B to soils. *Chemosphere*, *216*, 507–515. doi: 10.1016/j.chemosphere.2018.10.098.

Welch, B. L. (1951). ON THE COMPARISON OF SEVERAL MEAN VALUES: AN ALTERNATIVE APPROACH. *Biometrika*, *38*, 330–336. doi: 10.1093/biomet/38.3-4.330.

Williams, B., & Cremaschi, S. (2021). Selection of surrogate modeling techniques for surface approximation and surrogate-based optimization. *Chemical Engineering Research and Design*, *170*, 76–89. doi: 10.1016/j.cherd.2021.03.028.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259. doi: 10.1016/S0893-6080(05)80023-1.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*, 2839–2846. doi: 10.1016/j.patcog.2015.03.009.

Wu, G., Kechavarzi, C., Li, X., Wu, S., Pollard, S. J. T., Sui, H., & Coulon, F. (2013). Machine learning models for predicting PAHs bioavailability in compost amended soils. *Chemical Engineering Journal*, *223*, 747–754. doi: 10.1016/j.cej.2013.02.122.

Xiang, L., Xiao, T., Yu, P.-F., Zhao, H.-M., Mo, C.-H., Li, Y.-W., Li, H., Cai, Q.-Y., Zhou, D.-M., & Wong, M.-H. (2018). Mechanism and Implication of the Sorption of Perfluorooctanoic Acid by Varying Soil Size Fractions. *Journal of Agricultural and Food Chemistry*, *66*, 11569–11579. doi: 10.1021/acs.jafc.8b03492.

Xiao, F. (2017). Emerging poly- and perfluoroalkyl substances in the aquatic environment: A review of current literature. *Water Research*, *124*, 482–495. doi: 10.1016/j.watres.2017.07.024.

Xiao, F., Zhang, X., Penn, L., Gulliver, J. S., & Simcik, M. F. (2011). Effects of Monovalent Cations on the Competitive Adsorption of Perfluoroalkyl Acids by Kaolinite: Experimental Studies and Modeling. *Environmental Science & Technology*, *45*, 10028–10035. doi: 10.1021/es202524y.

Xie, G., Cui, J., Zhai, Z., & Zhang, J. (2020). Distribution characteristics of trifluoroacetic acid in the environments surrounding fluorochemical production plants in Jinan, China. *Environmental Science and Pollution Research*, *27*, 983–991. doi: 10.1007/s11356-019-06689-4.

Xie, J., Liu, S., Su, L., Zhao, X., Wang, Y., & Tan, F. (2024). Elucidating per- and polyfluoroalkyl substances (PFASs) soil-water partitioning behavior through explainable machine learning models. *Science of The Total Environment*, *954*, 176575. doi: 10.1016/j.scitotenv.2024.176575.

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, *2*, 249–262. doi: 10.1007/s41664-018-0068-2.

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*, 022022. doi: 10.1088/1742-6596/1168/2/022022.

Yuan, X., Ge, Z., & Song, Z. (2014). Soft sensor model development in multiphase/multimode processes based on Gaussian mixture regression. *Chemometrics and Intelligent Laboratory Systems*, *138*, 97–109. doi: 10.1016/j.chemolab.2014.07.013.

Zhang, X., Leng, S., Qiu, M., Ding, Y., Zhao, L., Ma, N., Sun, Y., Zheng, Z., Wang, S., Li, Y., & Guo, X. (2023). Chemical fingerprints and implicated cancer risks of Polycyclic aromatic hydrocarbons (PAHs) from fine particulate matter deposited in human lungs. *Environment International*, *173*, 107845. doi: 10.1016/j.envint.2023.107845.

Zhang, Z., Zhang, X., Zhang, D., Zhang, X., Qiu, F., Li, W., Liu, Z., Shu, J., & Tang, C. (2022). Application of Machine Learning in a Mineral Leaching Process Taking Pyrolusite Leaching as an Example. *ACS Omega*, *7*, 48130–48138. doi: 10.1021/acsomega.2c06129.

Zhu, J.-J., Yang, M., & Ren, Z. J. (2023). Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environmental Science & Technology*, acs.est.3c00026. doi: 10.1021/acs.est.3c00026.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*, 173–181. doi: 10.1348/000711004849222.

Zou, R., Lung, W.-S., & Wu, J. (2007). An adaptive neural network embedded genetic algorithm approach for inverse water quality modeling. *Water Resources Research*, *43*, 2006WR005158. doi: 10.1029/2006WR005158.