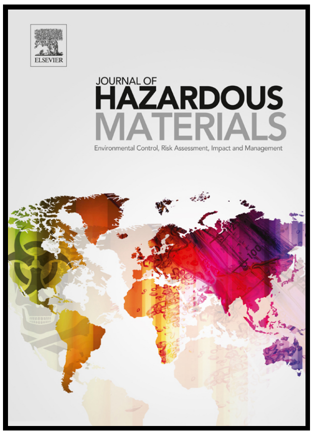


Use of a Large Dataset to Develop New Models for Estimating the Sorption of Active Pharmaceutical Ingredients in Soils and Sediments

Jun Li, John L. Wilkinson, Boxall

PII: S0304-3894(21)00652-X

DOI: https://doi.org/10.1016/j.jhazmat.2021.125688

Reference: HAZMAT125688

To appear in: *Journal of Hazardous Materials*

Received date: 25 January 2021
Revised date: 11 March 2021
Accepted date: 16 March 2021

Use of a Large Dataset to Develop New Models for Estimating the Sorption of Active Pharmaceutical Ingredients in Soils and Sediments

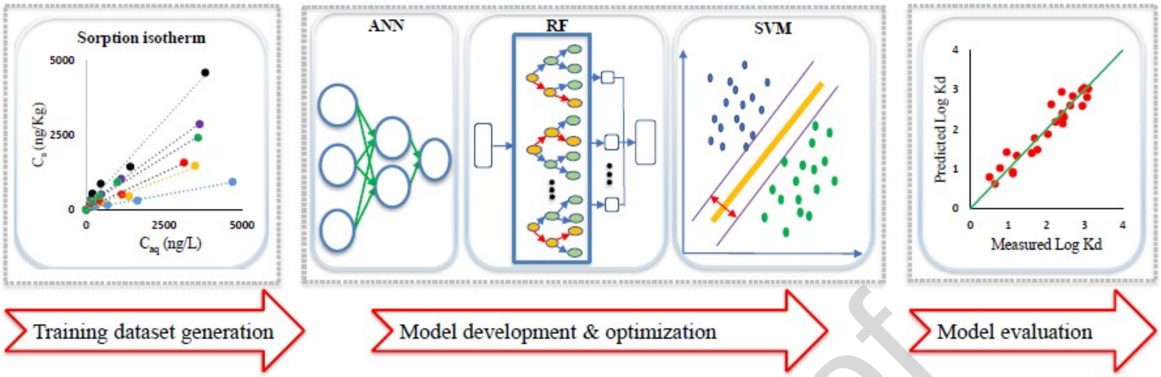
Jun Li, [†] **John L. Wilkinson,** [†] **and Alistair B.A. Boxall***[, †]

[†]**Department of Environment and Geography, University of York, Heslington, York, YO10 5NG, UK**

Abstract

Information on the sorption of active pharmaceutical ingredients (APIs) in soils and sediments is needed for assessing the environmental risks of these substances yet these data are unavailable for many APIs in use. Predictive models for estimating sorption could provide a solution. The performance of existing models is, however, often poor and most models do not account for the effects of soil/sediment properties which are known to significantly affect API sorption. Therefore, here, we use a high-quality dataset on the sorption behavior of 54 APIs in 13 soils and sediments to develop new models for estimating sorption coefficients for APIs in soils and sediments using three machine learning approaches (artificial neural network, random forest and support vector machine) and linear regression. A random forest-based model, with chemical and solid descriptors as the input, was the best performing model. Evaluation of this model using an independent sorption dataset from the literature showed that the model was able to predict sorption coefficients of 90% of the test set to within a factor of 10 of the experimental values. This new model could be invaluable in assessing the sorption behavior of molecules that have yet to be tested and in landscape-level risk assessments.

Graphical Abstract



Keywords:

Active pharmaceutical ingredients

Machine learning

Molecular and solid properties

Environmental fate

Environmental risk assessment

1. Introduction

Approximately 2100 different APIs are used to treat and prevent disease for humans and animals in the United Kingdom and new drugs are continually being developed (eMC, 2020; Burns *et al.,* 2018; Boxall *et al.,* 2003). A significant portion of APIs used in human medicine is discharged into wastewater treatment facilities via the excretion process or the disposal of unused or expired products and subsequently released into the environment via wastewater effluents, wastewater irrigation and sewage sludge application (Carter *et al.,* 2014; Jelic *et al.,* 2011). APIs used for veterinary purposes will be released to the environment by pasture animals, manure application or from aquaculture facilities. The occurrence of APIs in the environment has caused concerns throughout the world due to the potential for toxicological effects on aquatic and terrestrial organisms (aus der Beek *et al.,* 2016; Boxall, 2004).

An understanding of the sorption behavior of API in soils and sediments is important for environmental exposure assessment and for characterizing the environmental risks of a compound (Srinivasan *et al.,* 2014). Given the large number of APIs in use, modeling approaches that predict the sorption in soil and sediment are invaluable as they provide significant cost and time savings compared to laboratory-based methods (Barron *et al.,* 2009). Conventional modelling approaches typically use statistical linear regression analysis to develop a specific form of mathematical equation from empirical data to estimate the relationship between a dependent variable and one or more independent variables (Wu *et al.,* 2013). Using this traditional approach, it is hard to produce a universal form of equation to capture the highly non-linear relationships between the variables (Berthod *et al.,* 2017). An alternative is to use machine learning approaches which aim to identify patterns in data and use patterns to make predictions without being explicitly programmed (Wu *et al.,* 2013). Machine learning-based models have been reported to better map inputs and outputs efficiently in complex situations compared to traditional regression models (Giri *et al.,* 2011).

Over the past two decades, machine learning approaches (such as artificial neural networks, random forest and support vector regression etc.) have increasingly been used in the ecotoxicological field and applied to modelling the bioavailability or bioconcentration of organic compounds, where these approaches yielded good predictions with the overall accuracy (R^2_{test}) ranging from 0.725 to 0.954 (Miller *et al.,* 2019; Strempel *et al.,* 2013; Wu *et al.,* 2013; Zhao *et al.,* 2008; Liu *et al.,* 2006). To date, only a few attempts have been made to develop machine learning-based models for estimating sorption in sludge or soils (Berthod *et al.,* 2017; Barron *et al.,* 2009). For example, Barron *et al.* (2009) proposed an ANN model, that used 37 molecular descriptors as input variables, to estimate the sorption coefficients (Kd) of APIs in soil and sludge. Similar ANN models were developed using three sets of molecular descriptors (Molecular Operating Environment, VolSurf and ParaSurf) to predict Kd values for APIs in sludge with a minimum mean unsigned error on test sets (MUE_{test}) of 0.54 (Berthod *et al.,* 2017).

The currently available machine learning-based approaches were developed for single sludge and soil types. However, sorption coefficients of APIs are known to vary significantly depending on soil and sludge properties. For example, in recent experimental studies on the sorption of 21 APIs covering a range of physico-chemical properties and classes, the results showed that sorption coefficients across five soil types varied on average by a factor of 19 and by a maximum factor of 75 (Li *et al*, 2020). The size of training sets for some of the existing models is also limited (a maximum of 297 experimental Kd values) and none of the ANN-based models has undergone extensive evaluation against independent datasets. Work from other sectors shows that the availability of large datasets for model training is known to improve the predictive power of these approaches (Byvatov *et al.,* 2003). In addition to ANNs, other machine learning algorithms such as support vector machine and random forest are available and particularly powerful in describing the complex and non-linear relationships that exist between the input and output variables (Zhang *et al.,* 2020; Palmer *et al.,* 2007; Gao *et al.,* 1996). These approaches might offer an alternative to the ANN-based algorithms for modelling the sorption of APIs in soil and sediment.

The aim of this study was therefore to develop new machine learning-based models for estimating the sorption of ionisable (cationic, anionic and zwitterionic) and neutral APIs in both soil and sediments with varying properties. The specific objectives of the study were to: 1) develop a unique, large and high-quality dataset on the sorption coefficients of a diverse set of APIs in a range of soil and sediment types; 2) use the resulting data alongside a range of machine learning methods to develop new models for estimating sorption of APIs; and 3) evaluate the predictive ability of the developed models using two external validation datasets resampled from the literature.

2. Material and methods

2.1 Chemicals and Solvents

Fifty-eight APIs and deuterated forms of 28 of the APIs were purchased from Sigma-Aldrich (Gillingham, UK) (purity ≥95%). The APIs were selected based on their high usage, detection frequency in wastewater effluent, potential environmental concern and physicochemical properties. The study APIs covered 23 therapeutic classes including 35 bases, 8 acids, 7 neutral and 8 zwitterionic compounds at environmentally relevant pH values and covered wide hydrophobicity range ($-1.6 <$ Log Kow < 5.72). The CAS number, therapeutic class, and key physicochemical properties of the 58 APIs and associated deuterated compounds are provided in Table S1 (Appendix 1). Solvents including methanol, acetonitrile, dimethyl sulfoxide (DMSO) and water were LC-MS grade and were obtained from Fisher Scientific (Loughborough, UK). Ammonium formate (≥97%) and formic acid (≥95%) were purchased from Sigma-Aldrich (Gillingham, UK).

2.2 Soil and Sediment Samples

Previous studies have revealed that the sorption behavior of APIs is largely dependent on the sorbent physicochemical properties (Li *et al.*, 2020; Kodešová *et al.,* 2015). Therefore, in the present study, thirteen soils and sediments were selected to cover the variation in solid properties in the natural environment that may influence API sorption in soils and sediments in order to develop models that are broadly applicable across the landscape.

Five soils were obtained from LandLook (Midlands, UK) with three additional soils obtained from LUFA (Speyer, Germany). Five sediment samples were collected from the top 5cm sediment layer at five river and stream sites in North Yorkshire (UK, information on sample sites are provided in Table S2) using a sediment grab. All the samples were air-dried and sieved through a 2-mm mesh to ensure homogeneity and then stored in sterile sampling bags at 4℃ until use. To minimize biological activity, all the samples were sterilized by heating at 105℃ for 3 hours prior to use in the experiments. The characteristics of the soils and sediments were analyzed by Forest Research Company (Surrey, UK), and are summarized in Table S2. Further information on the measurement procedures for soil attributes are described by Li *et al*. (2020).

The soils and sediments represented general soil/sediment characteristics of European agricultural systems in terms of their pH (3.19-7.43), organic carbon content (1.34-5.90%), clay content (10-45%), and cation exchange capacity (4.42-25.11 cmol/kg). The key properties of the study soils and sediments were in good agreement with typical pH (3-8) and mean organic carbon content (<5%) ranges observed in the natural environment (Stockmann *et al.,* 2015; Ravisangar *et al.,* 2001; UK Soil Observatory Soils Map View).

2.3 Batch Sorption Experiments

Batch sorption experiments were based on the method described in OECD guideline 106 (OECD, 2000). A stock solution containing a mixture of 58 APIs at 20 mg/L was prepared in methanol and stored in the dark at -18℃ prior to the experiments. Preliminary experiments were performed to determine the equilibrium time and optimum solid-to-solution ratios for each soil and sediment (Table S3).

In the definitive sorption experiments, either 1 or 5 g soil or sediment was weighed into a Duran bottle (either 25, 50, 250 or 500 ml) and mixed with 0.01 M $CaCl_2$ solution (ranging from 10 to 400 ml) and shaken for 12 h to pre-equilibrate. Aliquots of the stock solution of the mixture of APIs were then spiked into the aqueous phase to yield initial concentrations of 10, 20, 30, and 40 μg/L. Triplicate samples were prepared for each solid-to-solution ratio and concentration. Control samples containing APIs in 0.01 M $CaCl_2$ solution (without soil or sediment) and blank samples containing soil or sediment mixed with $CaCl_2$ (without APIs) were prepared for each soil or sediment type. All the samples were agitated in the dark at 4℃ at 220 rpm for 24 h to reach equilibrium. Soil suspensions were centrifuged at 2000 rpm for 15 min and the supernatant was then taken and filtered through 0.45 μm glass fiber filters (Whatman). An aliquot of 0.2 mL of supernatant was then transferred into amber glass vials and diluted with 0.75 mL LC-MS grade water and 0.05 mL internal standard solution (containing 200 μg/L of each of the deuterated APIs in methanol) for determination of API concentrations by HPLC-MS/MS. A five times dilution was used to minimize the potential for $CaCl_2$ to damage the spectrometer and to reduce matrix effects.

Determination of the concentrations of APIs in the supernatant samples was achieved using a Thermo Scientific Dionex UltiMate 3000 HPLC coupled with a Thermo Scientific Endura TSQ triple-quadrupole mass spectrometer. The analytical method applied to determine the concentration of a mixture of 58 APIs was adapted from two existing methods (Wilkinson *et al.,* 2019; Furlong *et al.,* 2014). Details of the analytical method and method validation are provided in the Supporting Information.

2.4 Derivation of sorption coefficients

Linear, Freundlich and Langmuir isotherm models were fitted the batch sorption data. All the model parameters were evaluated by regression analysis using GraphPad Prism (version 7.00). The significance of the regressions and the correlation coefficient (R^2) were used to determine the best-fitting isotherm to the batch sorption data. The Linear, Freundlich and Langmuir isotherms were represented by Equations (2), (3), (4):

$$C_s = \frac{(C_i - C_{aq}) * V}{M} \tag{1}$$

$$C_s = K_d \cdot C_{aq} \tag{2}$$

$$\text{Log}\, C_s = \text{Log}\, K_f + \frac{1}{n} \text{Log}\, C_{aq} \tag{3}$$

$$C_{aq}/C_s = 1/(Q_m K_L) + C_{aq}/Q_m \tag{4}$$

Where C_i and C_{aq} are the initial and equilibrium concentration of the compound in aqueous solution (ng/L), V is the solution volume in the suspension (mL); M is the mass of soil (g); C_s is the concentration of API adsorbed in the sorbent phase at equilibrium (ng/kg). K_d, K_f and K_L are the Linear, Freundlich and Langmuir isotherm constants. n and Q_m are the Freundlich exponent and maximum Langmuir sorption capacity, respectively.

2.5 External Data Collection

Two external datasets (A and B) consisting of 583 linear sorption coefficients (Kd values) for 91 APIs (36 bases, 24 acids, 19 neutrals and 12 zwitterions) in soil or sediment, resampled from 30 published sorption studies, were collated to develop an independent dataset for testing the predictive capability and generalizability of the developed models (Table S11). The obtained sorption data were experimentally determined using the standard batch equilibrium sorption approach (OECD, 2000). These independent datasets comprise a broad range of API properties and solid characteristics representing the property space of APIs more generally and soil/sediment characteristics of global agricultural systems, with the associated sorption coefficients varying by seven orders of magnitude (0.05 < Kd < 1277873.9 L/kg). Dataset A contained 121 Kd values for 22 APIs with a complete record of soil or sediment characteristics including pH, TOC, texture, CEC as well as composition of exchangeable cations. The remaining data (462 Kd values) only provided standard soil or sediment properties (pH and TOC). The dataset B containing all 583 Kd values was applied to evaluate the performance of the models proposed based only on molecular descriptors.

2.6 Modelling Approaches

Artificial neural network (ANN), random forest (RF) and support vector machine (SVM) models as well as multiple linear regression model were developed, using the training sorption dataset (see Table S12) for estimating Kd values of pharmaceuticals from pharmaceutical descriptors and solid properties. ANNs usually comprise one input layer, one output layer as well as one or more hidden layers consisting of connected neurons to calculate the sum of input weights and produce the outcome through non-linear activation function (Gao *et al.,* 1996). RF is an ensemble learning method to achieve the final prediction by voting the prediction from decision trees using bootstrap aggregating algorithm (Han *et al.,* 2018). SVM is another non-linear algorithm that aims to find the optimal hyperplane that could separate the classes of data with the largest margin in a high-dimensional feature space (Souissi and Cherif, 2016).

Multilayer perceptrons (MLP, 3-5 layers) conducted in Python (python 3.7.3) with the MXNet Python package (version 1.5.0) were used to develop the ANN models. RF, SVM as well as linear regression were implemented in Python with the Scikit-learn package (version 0.21.2). The parameters of MLP including learning rate, epoch, the number of hidden layers and additional neurons as well as RF (Max-depth, min_sample_leaf, min_samle_split and N_estimators) and SVM (C and gamma) were set with an initial range and optimized by iteratively changing each parameter and repeatedly training in every iteration (the setting range of each parameter is shown in Table S9). Models of each machine learning approach were retrained more than 1000 times for parameter optimization. In order to avoid overfitting of the model, the internal prediction accuracy of each retrained model was evaluated using 10-fold cross-validation by resampling of the training subset. The root mean square error (RMSE) and coefficient of determination (R^2) were used to assess the prediction accuracy of the developed models. The parameters of optimal model with maximize accuracy (highest R^2 and lowest RMSE) were generated through the tree-structured parzen estimator approach (TPE) using Hyperopt package (version 0.1.2) in Python (Bergstra *et al.,* 2011).

Previous studies have assessed the importance of descriptors in the neural network models either by calculating the shift of model precision caused by removing one descriptor (Miller *et al.,* 2019; Miller *et al.,* 2016; Barron *et al.,* 2009) or through automatic relevance determination (Berthod *et al.,* 2017) or Pearson correlation analysis (Shi *et al.,* 2017). However, both ANN and SVM are regarded as the typical “black box” approaches that could produce fluctuating results in interpreting the contributions of individual variables in the model (Palmer *et al.,* 2007; Olden and Jackson, 2002). The reliability and stability of descriptor importance analysis of ANN and SVM approaches become limited when the model contains a large number of input variables, or where some variables display multicollinearity (Miller *et al.,* 2016; Burden *et al.,* 2000). The RF approach used in this study, on the other hand, provides a straightforward method for feature ranking by averaging the decrease of the weighted impurity of each feature (Nguyen *et al.,* 2015). The descriptor importance was determined in the RF approach using the Feature Importance algorithm implemented in Python with Scikit-Learn package (version 0.20.3). The predictive ability of the optimum models against the external datasets was further confirmed by Nash−Sutcliffe Efficiency (NSE). The approach for calculation of RMSE, NSE are described in our previous study (Li *et al*, 2020). To enable application of the developed models by end users, the Python code for the models is provided in GitHub repositories (https://github.com/Jun-Li-York/Sorption-model-code).

2.7 Feature Selection

Each machine leaning and linear regression approach was used to build two optimum models. One model was derived only from the molecular descriptors while the other model was developed also incorporating soil and sediment properties.

Two separate sets of input variables including 25 molecular and 13 solid descriptors that potentially influenced the API sorption were initially selected for the model development (Table S6). Molecular descriptors covering topological, constitutional, geometrical, physico-chemical properties as well as ionisation fraction were calculated using ACD-Labs (v5.0.0.184) and alvaDesc (v1.0.8) software using the canonical SMILES of the APIs as the input. These descriptors were then down-selected to reduce descriptor redundancy depending on the coefficient of variation (ratio of the standard deviation to the mean) and correlation coefficient of each variable with the other variables. A coefficient of variation (CV) of < 0.05 was used as cutoff value used to remove the variables with small variance (Table S7). The intercorrelation assessment of initial input variables is provided in Table S8. Strong intercorrelation ($r > 0.9$) was observed among a number of the variables (e.g., the significant correlation among molecular refractivity, molar volume, molecular weight and Ghose-Crippen molar refractivity; correlation between CEC and exchangeable Ca^{2+}, etc.), which may lead to unstable estimates of the model so only one variable was selected from each intercorrelated pair. The remaining descriptors, containing nineteen molecular and eight solid variables were generated into two optimal sets for the development of RF and ANN models (See Table S7). SVM initially did not perform well with the optimal sets and tend to overfit the training data. The input variables of SVM were then further down-selected using forward stepwise selection algorithms to remove redundant descriptors and minimize collinearity, which resulted in the selection of fourteen molecular and five solid variables (Table 1). Furthermore, collinearity diagnostics was performed to avoid multicollinearity in the linear regression model. The number of input variables was further reduced according to their variable inflation factor (VIF < 5), resulting in an optimum set of 6 molecular and 2 solid independent variables (Table S10).

3. Results and discussion

3.1 Sorption Behavior of Study Compounds

It was possible to develop sorption isotherm for 54 out of 58 study APIs in soil and sediment. For the other four compounds it was not possible to derive isotherms from the data due to either instability during the chemical analysis (cloxacillin), extensive adsorptive losses (> 80%) during the filtration process (miconazole) or an extremely high sorption affinity resulting in concentrations in the supernatant lower than limits of quantification (ketoconazole and itraconazole). The linear isotherm best described the sorption behaviour of the majority of the study compounds with 689 of the isotherms showing a statistically significant ($p < 0.05$) correlation between soil and water API concentrations, this was followed by the Freundlich isotherm (N = 596, $p < 0.05$) and then the Langmuir isotherm (N = 54, $p < 0.05$) (see Table S13). Therefore, the linear sorption coefficients were used for subsequent model development and are discussed below.

The Kd values of study APIs varied by five orders of magnitude across the different solid types with the lowest Kd (0.20 L/kg) being obtained for gabapentin and the highest Kd (35503 L/kg) being obtained for verapamil (Figure 1 and Table S13). Across the different classes of APIs, sorption coefficients for the basic and zwitterionic molecules varied by five orders of magnitude while sorption coefficients of the neutral and acidic molecules varied by three orders of magnitude suggesting that the sorption behavior of basic and zwitterionic APIs are more susceptible to differences in the properties of the solid matrices than acidic and neutral APIs. The sorption affinities of zwitterionic APIs (mean Kd of 2700 L/kg) and basic APIs (mean Kd of 657 L/kg) were generally higher than neutral (mean Kd of 30.4 L/kg) and acidic (mean Kd of 15.1 L/kg) APIs, suggesting charge forms and degree of ionisation at typical environmental pH of different classes of APIs are likely to drive the sorption of APIs in soil and sediment.

The sorption behaviour of APIs also displayed large variability within each study soil and sediment (Table S13). APIs exhibited higher sorption affinities to soil 8 and sediment 3 with the mean Kd values of 1358 and 1660 L/kg, respectively. Whereas lower sorption affinities of APIs were observed in soil 4 and sediment 4 with the mean Kd values of 219 and 486 L/kg being obtained, respectively. This observation revealed that the solid properties are also important in determining the API sorption in soil and sediment.

The present study used a high throughput approach to derive experimental Kd values involving testing 58 APIs in a multi-sorbate system over a narrow, but environmentally-relevant concentration range. We previously assessed 21 of the APIs in single sorbate studies at higher concentration levels (mg/L) in five of the test soils (Li *et al.,* 2020). Comparison of the Kd values measured in this study with values from our previous study showed very good agreement (R = 0.94, $p < 0.05$) indicating that our approach is robust. Sorption data are also available from the literature for many of the APIs with 27 and 23 out of 54 study APIs have been investigated in terms of their sorption behaviours in soil and sediment, respectively (Table S14). Comparison of our data with previous studies showed that the Kd values of APIs measured in test soils for bases (amitriptyline, atenolol, citalopram, erythromycin, fluoxetine, metformin, propranolol, tramadol, cimetidine, diazepam, metronidazole, trimethoprim and tylosin), acids (naproxen, sulfadiazine, sulfamethoxazole and warfarin), neutrals (caffeine, carbamazepine, paracetamol and temazepam) and zwitterions (ciprofloxacin and enrofloxacin) as well as the Kd values of APIs measured in test sediments for bases (atenolol, fluoxetine, propranolol, sitagliptin, tramadol, venlafaxine, cimetidine, diltiazem, lidocaine and trimethoprim), acids (naproxen, phenytoin and sulfamethoxazole), neutrals (fluconazole, paracetamol) and zwitterions (gabapentin, pregabalin) were in a similar range to sorption coefficients previously reported in the literature. However, our Kd values of ranitidine, phenytoin, oxytetracycline and tetracycline, measured in soils, and Kd values of amitriptyline, clarithromycin, metformin, verapamil, caffeine and carbamazepine, measured in sediments, were towards the higher end of the ranges previously reported (Table S14). This suggested that sorption coefficients of the majority of the study APIs generated from the multi-sorbate system used in the present study are generally consistent with previous research findings where studies have been done using single compounds and using different concentration ranges.

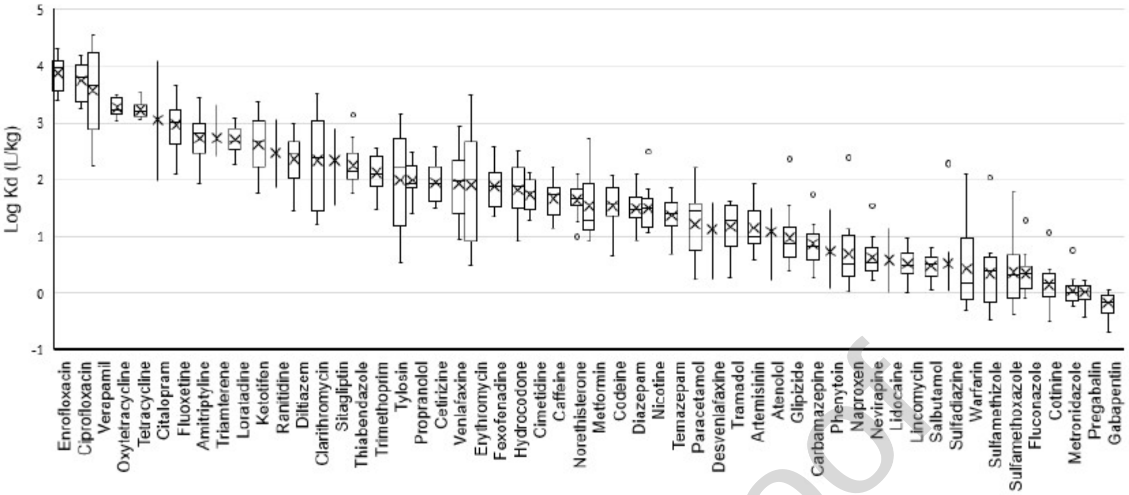


Figure 1 Overview of linear sorption coefficients for 54 study pharmaceuticals in 8 soils and 5 sediments. X markers=mean values; line through the box=median values; box=area between the 25th and the 75th percentile; whiskers=minimum and maximum values.

3.2 Overall Results of Proposed Models

The details of input variables, parameter optimization and statistical results of developed charge-specific models are shown in Table 1. Following parameter optimization, the six non-linear models developed using the RF, ANN and SVM approaches were well fitted to the measured Log Kd values with high precision (R^2_{train} > 0.889, $RMSE_{train}$ < 0.368). The machine learning-based models were superior to the two linear regression models (R^2_{train} < 0.583, $RMSE_{train}$ > 0.714) in capturing the sorption variance of APIs in soil and sediment by using corresponding shortlisted molecular and solid descriptors. All the proposed machine learning models exhibited higher R^2_{train} than those trained by Barron *et al.* (2009) (R^2_{train} =0.887) and Berthod *et al.* (2017) (R^2_{train} < 0.760), even though our experimental Log Kd values were measured in a diverse range of environmental solids. These results not only revealed that the measured sorption data generated from our standardized measurement procedure for training are highly reliable and consistent but also confirmed the appropriate selection of input variables for the model development and the advantage of TPE approach used in parameter optimization.

Based upon the 10-fold cross-validation results, the highest internal prediction accuracy ($RMSE_{CV}$ of 0.301) was obtained from the RF model using a combination of molecular descriptors and solid properties as input variables followed by ANN ($RMSE_{CV}$ of 0.380), SVN ($RMSE_{CV}$ of 0.471) and then linear regression ($RMSE_{CV}$ of 0.723). As expected, models incorporating solid descriptors and molecular descriptors generated from each approach consistently outperformed the corresponding model only relying on molecular descriptors. The significant improvements in R^2_{cv} by incorporating solid descriptors were observed for RF, ANN and linear regression models where the R^2_{cv} increased from 0.881 to 0.924, 0.820 to 0.878 and 0.521 to 0.562, respectively (Table 1). Whereas, inclusion of solid descriptors as SVM model inputs produced only a marginal improvement in the R^2_{cv} values (from 0.814 to 0.816). This result is in agreement with our previous finding that sorption models which take into account soil properties (such as exchangeable cations, TOC, clay content) yielded an improvement in the prediction of sorption for ionisable APIs in comparison to the models based only on physico-chemical properties of chemicals (Li *et al*, 2020).

Table 1 Results of parameter optimization and statistical analysis of the models developed to estimate the sorption of pharmaceuticals in soil and sediment ($N_{training}$ = 689).

Model	Descriptor	Parameter				Statistical result			
		Max_depth	**Min_sample_ leaf**	**Min_sample_split**	**N_etimators**	R^2_{train}	R^2_{cv}	$RMSE_{train}$	$RMSE_{cv}$
RF	Molecular (MV, UI, Charge 1^+, Log Sw, Log Dow, RB, Log Kow, Charge 1-, TPSA, *pKa*, Charge 0, HBA, HF, MlogP, DB, NR, Charge 1^+1^-, NSA, Nai), Solid (TOC, pH, ExK, ExNa, Silt, ExMg, CEC, Clay)	None	1	3	100	0.988	0.924	0.119	0.301
	Molecular (MV, UI, Charge 1^+, Log Sw, Log Dow,RB, Log Kow, Charge 1-, TPSA, *pKa*, Charge 0, HBA, HF, MlogP, DB, NR, Charge 1^+1^-, NSA, Nai)	None	1	6	100	0.959	0.881	0.225	0.378
	Descriptors	**Epoch**	**Learning rate**	**Additional neurons**	**Hidden layers**	R^2_{train}	R^2_{cv}	$RMSE_{train}$	$RMSE_{cv}$
ANN	Molecular (MV, UI, Charge 1^+, Log Sw, Log Dow, RB, Log Kow, Charge 1-, TPSA, *pKa*, Charge 0, HBA, HF, MlogP, DB, NR, Charge 1^+1^-, NSA, Nai), Solid (TOC, pH, ExK, ExNa, Silt, ExMg, CEC, Clay)	1000	0.0400970	12	3	0.980	0.878	0.156	0.380
	Molecular (MV, UI, Charge 1^+, Log Sw, Log Dow,RB, Log Kow, Charge 1-, TPSA, *pKa*, Charge 0, HBA, HF, MlogP, DB, NR, Charge 1^+1^-, NSA, Nai)	1000	0.0329548	15	3	0.911	0.820	0.329	0.463
	Descriptors	**C**	**Gamma**			R^2_{train}	R^2_{cv}	$RMSE_{train}$	$RMSE_{cv}$
SVM	Molecular (UI, Charge 1^+, Log Sw, Log Dow, Charge 1^-, TPSA, *pKa*, HBA, MlogP, DB, NR, Charge 1^+1^-, NSA, Nai), Solid (TOC, pH, ExK, ExNa, Silt)	19.6757274	0.0010360	NA	NA	0.935	0.816	0.281	0.471
	Molecular (UI, Charge 1^+, Log Sw, Log Dow, Charge 1^-, TPSA, *pKa*, HBA, MlogP, DB, NR, Charge 1^+1^-, NSA, Nai)	20.2757274	0.0010160	NA	NA	0.889	0.814	0.368	0.475
	Descriptors					R^2_{train}	R^2_{cv}	$RMSE_{train}$	$RMSE_{cv}$
Linear regression	Molecular (UI, Charge 1^+, Log Sw, Charge 1-, DB, Charge 1^+1^-), Solid (TOC, pH)	NA	NA	NA	NA	0.583	0.562	0.714	0.723
	Molecular (UI, Charge	NA	NA	NA	NA	0.539	0.521	0.750	0.756

	1^+, Log Sw, Charge 1-, DB, Charge 1^+1^-)								

Acronyms: MV, Molar volume; UI, Unsaturation index ; Charge 1^+, Charge fraction of plus 1; Log Sw, Aqueous solubility; Log Dow, Octanol/water partition coefficient corrected by soil/sediment pH; RB, Rotatable bonds; Log Kow, Octanol/water partition coefficient; Charge 1^-, Charge fraction of minus 1; TPSA, Fragment-based polar surface area from N, O, S, P polar coefficients; *pKa*, Acid/base dissociation constant; Charge 0, Charge fraction of neutral; HBA, Number of hydrogen Bond acceptors; HF, Hydrophilic factor; MlogP, Moriguchi log P; DB, Double bonds; NR, Number of rings; Charge 1^+1^-, Fraction of zwitterionic species; NSA, Number of sulfur atoms; Nai; number of hydrogens bound by the charged nitrogen; TOC, Total organic carbon content (%); pH, Soil or sediment pH; ExK, Exchangeable potassium (cmol/kg); ExNa, Exchangeable sodium (cmol/kg); Silt, Silt content (%); ExMg, Exchangeable magnesium (cmol/kg); CEC, Cation-exchange capacity (cmol/kg); Clay, Clay content (%).
Max_depth, N_etimators are the maximum depth of each tree and the number of trees in a random forest model, respectively.
Min_sample_leaf, Min_sample_split are the minimum number of samples required to be at a leaf node and the minimum number of samples required to split an internal leaf node, respectively.
C and Gamma are the regularization parameter and the relative weight of the regression error, respectively.
R^2_{train}, R^2_{cv} are R^2 on training set and cross-validated R^2, respectively.
$RMSE_{train}$, $RMSE_{CV}$ are root mean square error on training set and cross-validated root mean square error, respectively.

3.3 Relative Importance of Descriptors for Estimating Sorption Coefficients

Understanding the relative importance of individual descriptors to the models not only offers new insights into the mechanisms that drive the sorption of APIs in soil and sediment but also gives guidance on the selection of variables for future model development.

In the present study, molar volume (MV) was observed as the most important descriptors for estimating the sorption of APIs in soil and sediment (Figure 2). MV is defined as the molecular weight of a substance divided by its density, which is highly correlated to the molecular refractivity, molecular weight and, to some extent reflects the degree of hydrophobicity (see Table S8 for details on intercorrelation assessment). Previous research has shown that MV is an effective predictor employed in QSAR models for describing the sorption of negatively charged and uncharged APIs (Sathyamoorthy and Ramsburg, 2013). In addition to hydrophobic properties (MV, Log Sw, Log Dow, Log Kow) and unsaturation index, molecular charge fraction (Charge 1^+ and Charge 1^-) and dissociation constant (*pKa*) were top contributing descriptors in the RF model for estimating the sorption of APIs in soil and sediment, which supports the fact that the sorption behaviour of APIs to soils and sediment is charge-dependent (Li *et al*, 2020; Berthod *et al.,* 2017; Schaffer and Licha, 2015; Franco and Trapp, 2008). In comparison, the number of sulfur atoms, followed by the soil clay content and amine type of basic APIs make less of a contribution to the model (Figure 2).

Although previous studies have indicated that solid properties are important in driving the sorption behaviour of APIs (Li *et al*, 2020; Al-Khazrajy and Boxall, 2016; Kodešová *et al.,* 2015), in this study solid descriptors generally showed less influence on Log Kd prediction than molecular descriptors in the RF model. This could possibly be explained by the fact that the differences in sorption observed across APIs is much greater than that observed for a single API in a range of solid types. The most highly contributing solid descriptor was the TOC, suggesting that organic carbon might offer the principal sorption sites for hydrophobic sorption (van der Waals forces) as well as electrostatic sorption (ionic exchange, cation bridging, ligand exchange, and electron donor–acceptor interaction) of APIs in soil and sediment (Zhang *et al.,* 2017; Kodešová *et al.,* 2015). Soil/sediment pH also appears to play a role in driving the sorption behaviour of APIs due to the importance of the pH-adjusted lipophilicity (Log Dow) and molecular charge fraction as the model inputs. The diversity of the top contributing descriptors identified from descriptor importance analysis reflects the complexity of the interactions between APIs and soil/sediment.

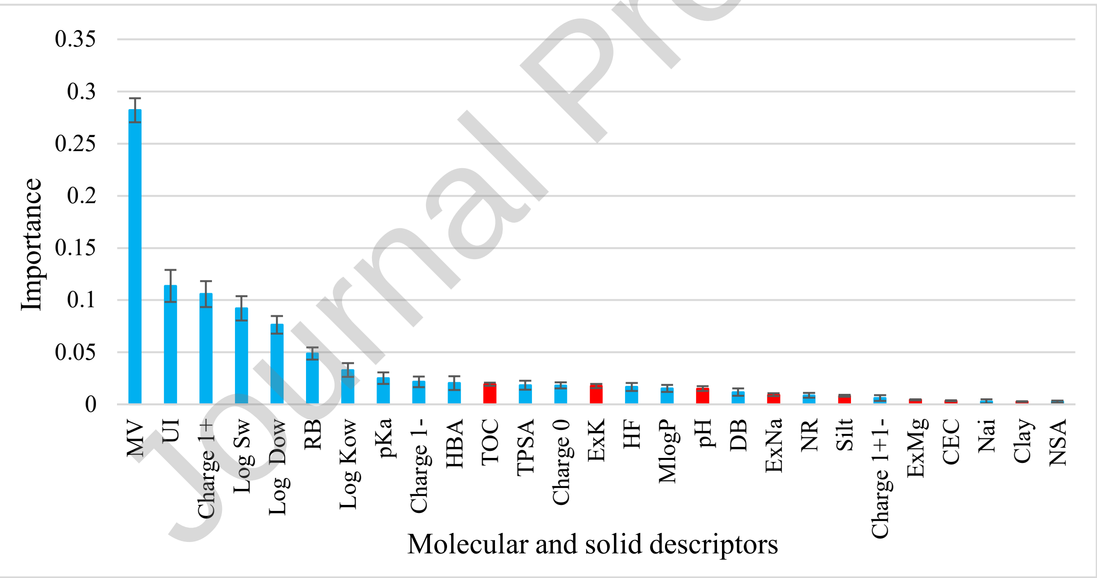


Figure 2 Relative importance of molecular (in blue) and solid (in red) descriptors to Random Forest model for estimating the sorption coefficient (Log Kd). Error bars represent standard error.

3.4 Model Performance Evaluation Based on External Datasets

The capability of the developed charge-specific models was assessed against two independent test datasets obtained from the literature (details in Table 2). The external datasets contained a diverse range of APIs and more than 63 % of these molecules had not been used to train our developed models. Comparison of predictions for these models with the experimentally-derived data showed that all of the developed models generated reasonable predictions of sorption behaviour (NSE > 0.302, $RMSE_{test}$ < SD). In the external dataset A, the RF model using molecular and solid descriptors achieved the best predictive performance ($RMSE_{test}$ of 0.636, Table 2), followed by ANN ($RMSE_{test}$ of 0.733), SVM ($RMSE_{test}$ of 0.772) and then linear regression ($RMSE_{test}$ of 0.774). Specifically, as shown in Table S15, the RF model was able to accurately describe the variability in sorption of APIs in both soil and sediment ($RMSE_{soil}$ of 0.598 and $RMSE_{sediment}$ of 0.697), with more than 90% of predicted Log Kd values being within a factor of 10 of the experimental values (N_{test} =121, see Figure 3A). The remaining three models also performed well at estimating sorption coefficients measured in soil ($RMSE_{soil}$ < 0.735, SD of 1.010), while these models appeared to slightly overestimate the sorption of basic APIs (e.g. amitriptyline, cimetidine, diltiazem and atenolol) over one order of magnitude in sediment ($RMSE_{sediment}$ > 0.815, SD of 0.750, Table S15). It should be noted that our solid descriptor set contained several soil/sediment properties (e.g. ExK, ExNa, ExMg, CEC), which are not commonly reported in literature studies, potentially limiting the broader applicability of the developed models. In the future, we recommend that researchers, investigating the sorption of ionisable compounds in soil, characterize test soils more extensively to aid in future model development.

In the external dataset B, according to their $RMSE_{test}$ values (Table 2), the three machine learning models achieved similar predictions and outperformed the linear regression models. The predicted data generated from the machine learning models derived only from molecular descriptors all generally agreed with the published sorption coefficients determined in soil and sediment, with more than 78% of the predictions estimated to be within 1 a factor of 10 of the corresponding observed values (N_{test} =583, $RMSE_{test}$ < 0.815, SD of 1.324 and NSE > 0.620, see Table 2). Specifically, these three models were able to predict Log Kd values relatively well for basic and neutral APIs in soil and sediment, while they all showed an underestimation of Log Kd values for zwitterionic APIs (e.g. norfloxacin, ciprofloxacin, danofloxacin and irbesartan) and overestimation for acidic APIs (e.g. gemfibrozil, bezafibrate, salicylic acid, sulfamethazine and indomethacin) by up to two orders of magnitude (Figure 3E, F and G). This suggests that solid properties play an important role in driving the sorption models for zwitterionic and acidic APIs as they could be contributing to additional sorption mechanisms including hydrophobic partitioning to soil organic carbon and electrostatic interactions among the charged species, electronegative solid surfaces (clay or organic matter) and soil exchangeable cations (Klement *et al.,* 2018; Kodešová *et al.,* 2015; Estevez *et al.,* 2014). Moreover, the relatively small size of the sorption data for model training was obtained for zwitterions (N=91) and acids (N=89) compared to bases (N=411) due to analytical limitations and the experimental conditions, which might result in poorer generalizability to unknown molecules. A large and high-quality sorption data for zwitterionic and acidic APIs obtained from the standardization of experimental procedures and analytical methods is highly warranted, which might improve the predictive performance of future models. A poorer performance was observed with the linear regression model, with a larger average error ($RMSE_{test}$ of 0.935, Table 2). This suggests that complex interactions involved in multiple mechanisms (such as hydrophobic forces as well as electrostatic interactions) exist between the sorbent and sorbate, linear regression modelling is less suited than machine learning approaches to dealing with this type of complexity.

Table 2 Comparison of model performance against the external sorption data.

Model	**External data set**	**Descriptor**	**N**	**SD**	R^2_{test}	$RMSE_{test}$	NSE_{test}
RF	A	19 Molecular + 8 Solid	121	0.931	0.695	0.636	0.529
	B	19 Molecular	583	1.324	0.644	0.815	0.620
ANN	A	19 Molecular + 8 Solid	121	0.931	0.607	0.733	0.375
	B	19 Molecular	583	1.324	0.680	0.787	0.646
SVM	A	13 Molecular + 5 Solid	121	0.931	0.521	0.772	0.307
	B	13 Molecular	583	1.324	0.635	0.812	0.623
Linear regression	A	6 Molecular + 2 Solid	121	0.931	0.624	0.774	0.302
	B	6 Molecular	583	1.324	0.514	0.935	0.500

N is the number of the observations. SD is the standard deviation of the observations. RMSE is the root mean square error. NSE is the Nash−Sutcliffe Efficiency value.

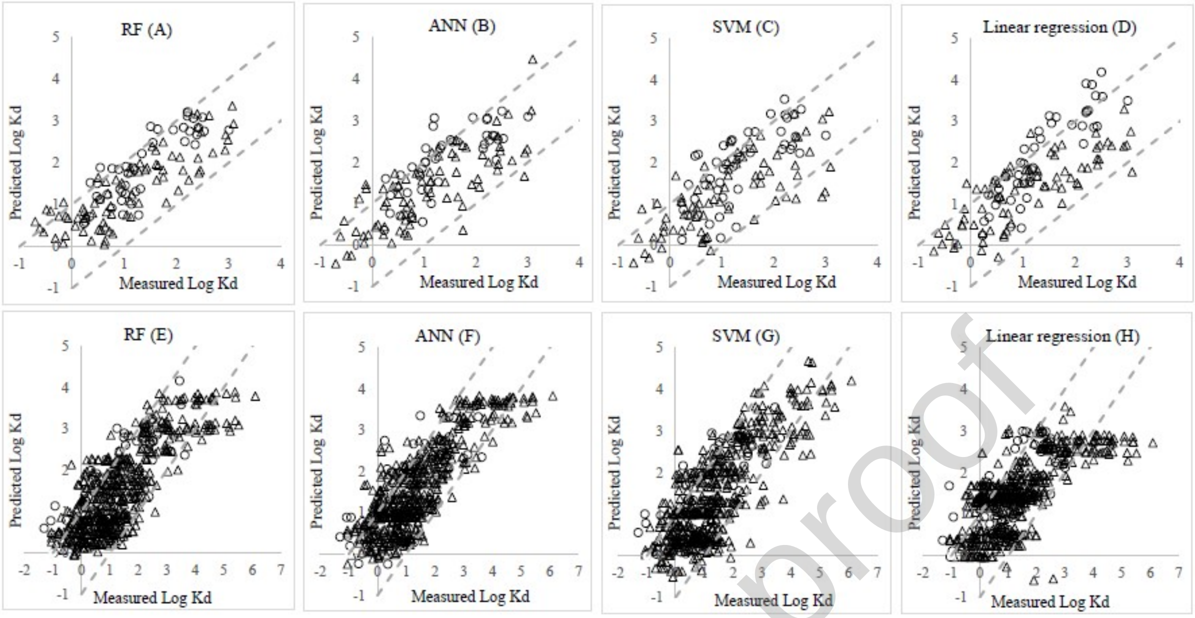


Figure 3 Comparison of predictive performance of (A) random forest, (B) artificial neural networks, (C) support vector machine and (D) linear regression models against the external sorption dataset A (N=121) and predictive performance of (E) random forest, (F) artificial neural networks, (G) support vector machine and (H) linear regression models against the external sorption dataset B (N=583). The dashed lines correspond to the predicted sorption coefficients ± 1 units against the measured values collected from literature. Open triangles and circles represent the Log Kd values observed in soil and sediment environments, respectively.

4. Conclusion

Our results demonstrate the power of using non-linear approaches for estimating the Log Kd from molecular and solid descriptors. Among the three-machine learning and regression approaches, the random forest method was found to be superior to other methods in both qualitative (in that it offered some mechanistic understanding from its feature-ranking function) and quantitative performance. In addition to the satisfactory general predictive performance, random forest offered many attractive advantages over the other methods as it: is immune to overfitting, generates an internal assessment of descriptor importance, has high learning speed, is insensitive to scaling of features and has good stability (Han *et al.,* 2018; Palmer *et al.,* 2007). Therefore, the random forest method could be used as a rapid and powerful tool for predicting the sorption of APIs in both soil and sediment for use in risk assessment of APIs and has the greatest potential to be applied to other endpoints relevant to environmental risk assessment or to modelling these endpoints for other environmental contaminants such as pesticides, polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs) and aliphatic hydrocarbons. For practical purposes, such an approach could considerably improve and harmonize the current risk assessment practices for estimating soil/sediment sorption in the REACH (European Commission, 2006) and EMEA (European Medicines Agency, 2006) guidelines for environmental risk assessment, and provide a rapid solution for regulators to support environmental decision making.

Supporting Information Description

Detailed information on study APIs, soils and sediments, preliminary sorption experiments, analytical methods, feature selection and parameter optimization, sorption isotherms as well as details of training and external evaluation data sets and model evaluation results.

CRediT authorship contribution statement

Jun Li: Conceptualization, Methodology, Formal analysis, Investigation, Software, Visualization, Writing - original draft. John L. Wilkinson: Methodology, Writing - review & editing. Alistair B.A. Boxall: Supervision, Conceptualization, Methodology, Writing - review & editing, Funding acquisition.

Funding

The study was funded by the EU/EFPiA Innovative Medicines Initiative Joint Undertaking under iPiE grant agreement (Number: 115735).

Corresponding Author

*E-mail: alistair.boxall@york.ac.uk; Tel: +44 (0)1904 324791; fax: +44 (0)1904 322998.

Notes

The authors declare no financial conflict of interest.

Acknowledgments

We would like to thank Edmund Bergstrom (Centre of Excellence in Mass Spectrometry, university of York) for his assistance in analytical method development. We also thank Hanting Xie (Data scientist at Amazon Web Services) for his valuable contribution to the parameter optimization and coding.

References

Al-Khazrajy, O. S., & Boxall, A. B. (2016). Impacts of compound properties and sediment characteristics on the sorption behaviour of pharmaceuticals in aquatic systems. *Journal of hazardous materials*, *317*, 198-209.

aus der Beek, T., Weber, F. A., Bergmann, A., Hickmann, S., Ebert, I., Hein, A., & Küster, A. (2016). Pharmaceuticals in the environment—Global occurrences and perspectives. *Environmental toxicology and chemistry*, *35*(4), 823-835.

Barron, L., Havel, J., Purcell, M., Szpak, M., Kelleher, B., & Paull, B. (2009). Predicting sorption of pharmaceuticals and personal care products onto soil and digested sludge using artificial neural networks. *Analyst*, *134*(4), 663-670

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems* (pp. 2546-2554).

Berthod, L., Whitley, D. C., Roberts, G., Sharpe, A., Greenwood, R., & Mills, G. A. (2017). Quantitative structure-property relationships for predicting sorption of pharmaceuticals to sewage sludge during waste water treatment processes. *Science of the Total Environment*, *579*, 1512-1520.

Boxall, A. B., Kolpin, D. W., Halling-Sørensen, B., & Tolls, J. (2003). Peer reviewed: are veterinary medicines causing environmental risks? *Environmental science & technology*, *37*(15), 286A-294A.

Boxall, A. B. (2004). The environmental side effects of medication: How are human and veterinary medicines in soils and water bodies affecting human and environmental health?. *EMBO reports*, *5*(12), 1110-1116.

Burden, F. R., Ford, M. G., Whitley, D. C., & Winkler, D. A. (2000). Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *Journal of Chemical Information and Computer Sciences*, *40*(6), 1423-1430.

Burns, E. E., Carter, L. J., Snape, J., Thomas-Oates, J., & Boxall, A. B. (2018). Application of prioritization approaches to optimize environmental monitoring and testing of pharmaceuticals. *Journal of Toxicology and Environmental Health, Part B*, *21*(3), 115-141.

Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences*, *43*(6), 1882-1889.

Carter, L. J., Garman, C. D., Ryan, J., Dowle, A., Bergström, E., Thomas-Oates, J., & Boxall, A. B. (2014). Fate and uptake of pharmaceuticals in soil–earthworm systems. *Environmental science & technology*, *48*(10), 5955-5963.

Datapharm Communications Limited. 2020. The Electronic Medicines Compendium (eMC). *Medicines.* Accessed August 17, 2020. https://www.medicines.org.uk/emc/browse

EMEA (European Medicines Agency). (2006). *Guideline on the Environmental Risk Assessment of Medicinal Products for Human Use*, EMEA/CHMP/SWP/4447/00.

Estevez, E., Hernandez-Moreno, J. M., Fernandez-Vera, J. R., & Palacios-Diaz, M. P. (2014). Ibuprofen adsorption in four agricultural volcanic soils. *Science of the Total Environment*, *468*, 406-414.

European Commission (2006). *Registration, Evaluation, Authorization and Restriction of Chemicals (REACH)*. Regulation (EC) No. 1907/2006 of the European Parliament and of the Council.

Furlong, E. T., Noriega, M. C., Kanagy, C. J., Kanagy, L. K., Coffey, L. J., & Burkhardt, M. R. (2014). Determination of human-use pharmaceuticals in filtered water by direct aqueous injection—high-performance liquid chromatography/tandem mass spectrometry. *US Geological Survey Techniques and Methods*, *5*, 49.

Franco A., & Trapp S. (2008). Estimation of the soil-water partition coefficient normalized to organic carbon for ionisable organic chemicals. Environmental Toxicology and Chemistry 27 (10): 1995–2004.

Gao, C., Govind, R., & Tabak, H. H. (1996). Predicting soil sorption coefficients of organic chemicals using a neural network model. *Environmental Toxicology and Chemistry: An International Journal*, *15*(7), 1089-1096.

Giri, A. K., Patel, R. K., & Mahapatra, S. S. (2011). Artificial neural network (ANN) approach for modelling of arsenic (III) biosorption from aqueous solution by living cells of Bacillus cereus biomass. *Chemical Engineering Journal*, *178*, 15-25.

Han, T., Jiang, D., Zhao, Q., Wang, L., & Yin, K. (2018). Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control*, *40*(8), 2681-2693.

Nguyen, T. T., Huang, J. Z., & Nguyen, T. T. (2015). Unbiased feature selection in learning random forests for high-dimensional data. *The Scientific World Journal*, *2015*.

Jelic, A., Gros, M., Ginebreda, A., Cespedes-Sánchez, R., Ventura, F., Petrovic, M., & Barcelo, D. (2011). Occurrence, partition and removal of pharmaceuticals in sewage water and sludge during wastewater treatment. *Water research*, *45*(3), 1165-1176.

Klement, A., Kodešová, R., Bauerová, M., Golovko, O., Kočárek, M., Fér, M., ... & Grabic, R. (2018). Sorption of citalopram, irbesartan and fexofenadine in soils: Estimation of sorption coefficients from soil properties. *Chemosphere*, *195*, 615-623.

Kodešová, R., Grabic, R., Kočárek, M., Klement, A., Golovko, O., Fér, M., ... & Jakšík, O. (2015). Pharmaceuticals' sorptions relative to properties of thirteen different soils. *Science of the total environment*, *511*, 435-443.

Li, J., Carter, L. J., & Boxall, A. B. (2020). Evaluation and development of models for estimating the sorption behaviour of pharmaceuticals in soils. *Journal of hazardous materials*, 122469.

Liu, H., Yao, X., Zhang, R., Liu, M., Hu, Z., & Fan, B. (2006). The accurate QSPR models to predict the bioconcentration factors of nonionic organic compounds based on the heuristic method and support vector machine. *Chemosphere*, *63*(5), 722-733.

Miller, T. H., Baz-Lomba, J. A., Harman, C., Reid, M. J., Owen, S. F., Bury, N. R., ... & Barron, L. P. (2016). The first attempt at non-linear in silico prediction of sampling rates for polar organic chemical integrative samplers (POCIS). *Environmental science & technology*, *50*(15), 7973-7981.

Miller, T. H., Gallidabino, M. D., MacRae, J. I., Owen, S. F., Bury, N. R., & Barron, L. P. (2019). Prediction of bioconcentration factors in fish and invertebrates using machine learning. *Science of the Total Environment*, *648*, 80-89.

OECD Guidelines for the Testing of Chemicals: Test No. 106 Adsorption Desorption Using a Batch Equilibrium Method (2000); Organization for Economic Cooperation and Development: Paris, France. Http: www.oecd.org/env/ehs/testing/TG_List_EN_Jul_2013.pdf.

Olden, J. D., & Jackson, D. A. (2002). Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, *154*(1-2), 135-150.

Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. (2007). Random forest models to predict aqueous solubility. *Journal of chemical information and modeling*, *47*(1), 150-158.

Ravisangar, V., Dennett, K. E., Sturm, T. W., & Amirtharajah, A. (2001). Effect of sediment pH on resuspension of kaolinite sediments. *Journal of environmental engineering*, *127*(6), 531-538.

Sathyamoorthy, S., & Ramsburg, C. A. (2013). Assessment of quantitative structural property relationships for prediction of pharmaceutical sorption during biological wastewater treatment. *Chemosphere*, *92*(6), 639-646.

Schaffer, M., & Licha, T. (2015). A framework for assessing the retardation of organic molecules in groundwater: Implications of the species distribution for the sorption-influenced transport. *Science of the Total Environment*, *524*, 187-194.

Shi, X., Tian, S., Yu, L., Li, L., & Gao, S. (2017). Prediction of soil adsorption coefficient based on deep recursive neural network. *Automatic Control and Computer Sciences*, *51*(5), 321-330.

Souissi, N., & Cherif, A. (2016). Artificial neural networks and support vector machine for voice disorders identification. *Int. J. Adv. Comput. Sci. Appl.*, *7*(5), 339-344.

Srinivasan, P., Sarmah, A. K., & Manley-Harris, M. (2014). Sorption of selected veterinary antibiotics onto dairy farming soils of contrasting nature. *Science of the Total Environment*, *472*, 695-703.

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., ... & Field, D. J. (2015). Global soil organic carbon assessment. *Global Food Security*, *6*, 9-16.

Strempel, S., Nendza, M., Scheringer, M., & Hungerbühler, K. (2013). Using conditional inference trees and random forests to predict the bioaccumulation potential of organic chemicals. *Environmental toxicology and chemistry*, *32*(5), 1187-1195.

UKSO. UK Soil Observatory Soils Map View. http://mapapps2.bgs.ac.uk/uk/ukso/home.html.

Wilkinson, J. L., Boxall, A., & Kolpin, D. W. (2019). A novel method to characterise levels of pharmaceutical pollution in large-scale aquatic monitoring campaigns. *Applied Sciences*, *9*(7), 1368.

Wu, G., Kechavarzi, C., Li, X., Wu, S., Pollard, S. J., Sui, H., & Coulon, F. (2013). Machine learning models for predicting PAHs bioavailability in compost amended soils. *Chemical engineering journal*, *223*, 747-754.

Zhang, Y., Price, G. W., Jamieson, R., Burton, D., & Khosravi, K. (2017). Sorption and desorption of selected non-steroidal anti-inflammatory drugs in an agricultural loam-textured soil. *Chemosphere*, *174*, 628-637

Zhang, K., Zhong, S., & Zhang, H. J. (2020). Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. *Environmental Science & Technology*.

Zhao, C., Boriani, E., Chana, A., Roncaglioni, A., & Benfenati, E. (2008). A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere*, *73*(11), 1701-1707.

Author Credit Statement

Jun Li: Conceptualization, Methodology, Formal analysis, Investigation, Software, Visualization, Writing - original draft. John L. Wilkinson: Methodology, Writing - review & editing. Alistair B.A. Boxall: Supervision, Conceptualization, Methodology, Writing - review & editing, Funding acquisition.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights

1. Linear sorption coefficients were generated for 689 pharmaceutical/substrate combinations.

2. A random forest model achieved excellent performance for estimating sorption of pharmaceuticals.

3. The new model provides a valuable tool for environmental risk assessment of pharmaceuticals