Environmental Science
Water Research & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Water Res. Technol.*, 2025, **11**, 481

A machine learning framework to predict PPCP removal through various wastewater and water reuse treatment trains†

Joung Min Choi, ‡[a] Vineeth Manthapuri, ‡[b] Ishi Keenum,[bc] Connor L. Brown,[d] Kang Xia,[e] Chaoqi Chen,[e] Peter J. Vikesland,[b] Matthew F. Blair,[b] Charles Bott,[f] Amy Pruden *[b] and Liqing Zhang*[a]

The persistence of pharmaceuticals and personal care products (PPCPs) through wastewater treatment and resulting contamination of aquatic environments and drinking water is a pervasive concern, necessitating means of identifying effective treatment strategies for PPCP removal. In this study, we employed machine learning (ML) models to classify 149 PPCPs based on their chemical properties and predict their removal *via* wastewater and water reuse treatment trains. We evaluated two distinct clustering approaches: C1 (clustering based on the most efficient individual treatment process) and C2 (clustering based on the removal pattern of PPCPs across treatments). For this, we grouped PPCPs based on their relative abundances by comparing peak areas measured *via* non-target profiling using ultra-performance liquid chromatography-tandem mass spectrometry through two field-scale treatment trains. The resulting clusters were then classified using Abraham descriptors and $\log K_{ow}$ as input to the three ML models: support vector machines (SVM), logistic regression, and random forest (RF). SVM achieved the highest accuracy, 79.1%, in predicting PPCP removal. Notably, a 58–75% overlap was observed between the ML clusters of PPCPs and the Abraham descriptor and $\log K_{ow}$ clusters of PPCPs, indicating the potential of using Abraham descriptors and $\log K_{ow}$ to predict the fate of PPCPs through various treatment trains. Given the myriad of PPCPs of concern, this approach can supplement information gathered from experimental testing to help optimize the design of wastewater and water reuse treatment trains for PPCP removal.

Received 2nd November 2024,
Accepted 18th December 2024

DOI: 10.1039/d4ew00892h

rsc.li/es-water

Water impact

Here we introduce a machine learning approach to predict the removal of pharmaceuticals and personal care products (PPCPs) during wastewater and water reuse treatment. By reducing the need for costly and labor-intensive analytical testing, this approach supports assessment of treatment efficacy and optimization treatment processes for efficient removal of various PPCPs.

1. Introduction

Pharmaceuticals and personal care products (PPCPs), comprising over 4000 diverse natural and synthetic substances, are widely used in medicine, industry, and consumer products.[1–3] Their extensive use has led to widespread occurrence in water bodies, as conventional wastewater treatment plants are not specifically designed for their removal.[4–6] This has raised concerns in various water reuse scenarios, including irrigation, groundwater recharge, and indirect potable reuse.[7,8] The presence of PPCPs in aquatic environments poses ecotoxicological risks, including endocrine disruption and potential human health hazards, even at trace concentrations.[9–12] These concerns are particularly relevant in water reuse contexts, where there is increased potential for human exposure.[13,14]

The chemical diversity of PPCPs and their typically low concentrations pose significant challenges to identifying effective removal processes.[15–17] The removal of PPCPs during

[a] *Department of Computer Science, Virginia Tech, Blacksburg, VA, 24061, USA. E-mail: lqzhang@cs.vt.edu*
[b] *Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, 24061, USA. E-mail: apruden@vt.edu*
[c] *Civil, Environmental and Geospatial Engineering, Michigan Tech University, MI, 49931, USA*
[d] *Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, 24061, USA*
[e] *School of Plant and Environmental Sciences, Blacksburg, VA, 24061, USA*
[f] *Hampton Roads Sanitation District, Virginia Beach, VA, 23455, USA*
† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4ew00892h
‡ These authors contributed equally to this work.

treatment is influenced by both operational factors and the inherent chemical properties of the compounds, such as molecular weight, hydrophobicity, and charge, which ultimately dictate their fate and removal efficiencies *via* various treatment processes.[18–23]

Abraham descriptors have recently been proposed to categorize the general chemical properties of PPCPs[24–27] and thus could aid in predicting their removal *via* various treatment processes. These descriptors offer a comprehensive profile of a compound's solvation properties and molecular interactions, which directly relate to whether a PPCP is likely to be removed *via* biological, sorptive, or oxidative processes, *etc.*[28] By quantifying the molecular interactions that govern sorption processes, Abraham descriptors could provide a systematic framework for understanding and predicting PPCP behavior in various treatment systems.

The advancement of analytical technologies for PPCP detection first brought to light their widespread occurrence in aquatic environments and remain the gold standard for measuring their removal *via* various treatment processes. Over the past two decades, early detection methods like gas chromatography/mass spectrometry (GC/MS) have evolved into non-targeted monitoring approaches such as ultra-performance liquid chromatography tandem mass spectrometry (UPLC-MS/MS).[29–31] These advanced methods enable the simultaneous analysis of multiple PPCPs with diverse properties, offering a comparative evaluation of removal efficiencies across a wide array of compounds. Conventional monitoring of PPCPs through treatment trains remains essential for regulatory compliance and watershed management. Field measurements provide direct evidence of PPCP occurrence and removal, but are resource-intensive, requiring specialized expertise and costly instrumentation.[32–35] The increasing complexity of data yielded by advanced analytical technologies like UPLC-MS/MS further complicates interpretation, underscoring the need for approaches to fully leverage this wealth of information. Analytical methods are particularly insufficient for emerging contaminants, where methods are not yet available, or if the aim is to assess a broad spectrum of PPCPs of concern.[36,37]

Recently, machine learning (ML) has shown promise in water and wastewater treatment studies,[38,39] providing a powerful array of tools for revealing important patterns in complex data sets, including non-linear relationships.[40,41] However, previous works have employed single ML frameworks, which incurs some drawbacks.[42–44] Supervised methods alone often struggle with high-dimensional, complex data and may miss important underlying patterns, while unsupervised techniques in isolation lack predictive capabilities. Given the intricate nature of PPCP behavior across various treatment processes, a more comprehensive approach is needed.

In this study, we propose a novel two-step ML approach to characterize and predict PPCP removal in wastewater and water reuse treatment systems. The method begins with unsupervised learning (clustering) to uncover inherent patterns and reduce dimensionality of complex datasets. This crucial step reveals natural groupings of PPCPs without relying on predefined labels. We then leverage these insights in a supervised ML classification phase, establishing quantitative relationships between PPCP properties and removal outcomes. This combined unsupervised-supervised approach is particularly well-suited to complex datasets, revealing otherwise hidden patterns and providing predictive capacity. By addressing the limitations of single-algorithm methods, our framework could provide a more robust and nuanced understanding of PPCP fate during wastewater and reuse treatment. The approach is demonstrated utilizing UPLC-MS/MS data from two full-scale wastewater treatment facilities that are respectively followed by a series of treatments for non-potable and potable reuse.

2. Materials and methods

This study was conducted in four steps: (1) sampling, analysis, and data collection (2) clustering PPCPs based on their relative abundances through various stages of wastewater and water reuse treatment (3) classification of PPCPs in each cluster based on the Abraham descriptors and $\log K_{ow}$ values and (4) validation using cross validation and statistical testing. The workflow is shown in Fig. 1.

2.1 Data collection

2.1.1 Facilities and sample collection. Two full-scale facilities employing activated sludge wastewater treatment followed by distinct water reuse treatments were the subject of this study. A non-potable treatment plant employed

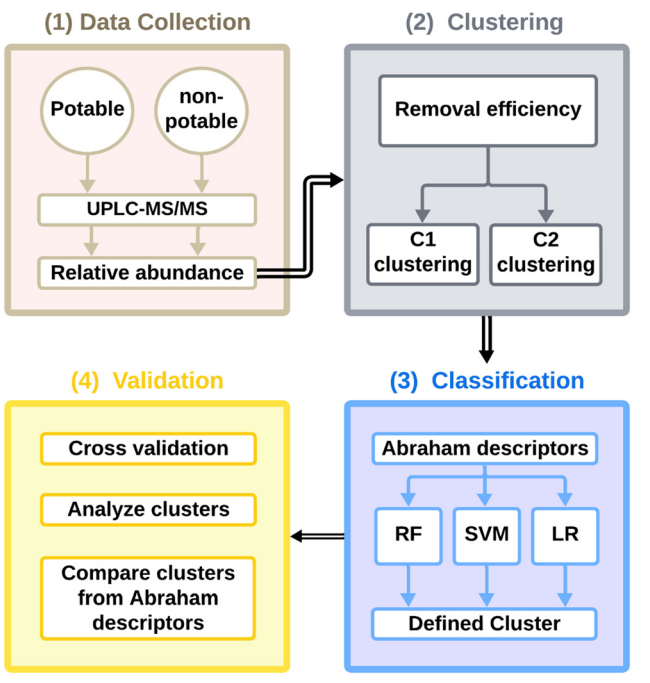


Fig. 1 Approach applied to characterize the removal of PPCPs through various wastewater and water reuse treatment processes and to develop the ML framework (RF = random forest, SVM = support vector machines, LR = logistic regression).

denitrification-filtration/chlorination prior to distribution of reused water primarily used for irrigation. An indirect potable reuse plant employed advanced water treatment (FlocSed/Ozone/BAC/GAC/UV) prior to aquifer recharge. In total, 84 samples were collected from the two treatment facilities between November 2018 and August 2019, as described in a companion study focused on the fate of antibiotic resistant bacteria and antibiotic resistance genes.[45]

2.1.2 PPCP multi-compound screening. All samples were processed immediately upon receipt and underwent pre-filtration using 0.7 μm glass fiber filters (Whatman, Maidstone, UK) and subdivided into 200 mL triplicates for analysis. As detailed in Section S1 of the ESI,† analytes were extracted, background matrices were cleaned up, and the final concentrates were obtained through solid-phase extraction (SPE). The resulting extracts were screened for the presence of 149 PPCPs, metabolites, dietary substances, agricultural chemicals, illegal drugs, and drug-testing agents that are commonly encountered in wastewater using UPLC/MS/MS (broadly referred to as "PPCPs" in this study), employing a semi-quantitative approach with a custom-made compound identification database.[46,47] For the SPE, Oasis HLB cartridges from Waters with a 60 mg sorbent bed mass and 3 mL reservoir volume were used. The cartridges were pre-conditioned with 3 mL HPLC-grade methanol and 3 mL ultra-pure water. Subsequently, samples were processed through the cartridges at 5 mL min^{-1}, and analytes were eluted with 3 mL HPLC-grade methanol, dried under N_2 gas on a vacuum evaporation system (Labconco Kansas City, MO), and reconstituted with 1 mL HPLC-grade acetonitrile–water solution (1 : 1, v : v). UPLC-MS/MS was conducted using a 1290 UPLC/Agilent 6490 Triple Quad tandem MS (Agilent Technologies Inc., Santa Clara, CA). All samples were processed, cleaned, and analyzed within a single analytical batch to support comparisons of relative differences by comparing peak areas across samples.

2.1.3 Abraham descriptors and $\log K_{ow}$ values. Abraham descriptors are parameters used to quantify key properties of a given chemical that govern its amenability to solvation and sorption. These descriptors include E (polarizability), S (dipolarity), A and B (hydrogen bond donating and accepting potential), V (molecular volume), and L (gas–hexadecane partition coefficient) (Table 1). Together, these parameters characterize the solvation properties of a compound: E and S relate to cavitation and van der Waals interactions; A and B characterize solute–solvent hydrogen bonding; V determines molecular size compatibility with substrate gaps; and L depicts bulk transport.

Table 1 Abraham descriptors for PPCP analysis

S no.	Symbol	Chemical characteristic	Units
1	E	Excess molar refraction	$cm^3\ mol^{-1}/10$
2	S	Dipolarity/polarizability	Dimensionless
3	A	Hydrogen bonding acidity	Dimensionless
4	B	Hydrogen bonding basicity	Dimensionless
5	V	McGowan characteristic volume	$cm^3\ mol^{-1}/100$
6	L	Gas-to-hexadecane partition	Dimensionless

We obtained the six Abraham descriptors for each PPCP from the publicly-available Helmholtz Centre for Environmental Research-Linear Solvation Energy Relationship (UFZ-LSER) database.[48] This database provides experimental Abraham descriptors for most compounds using compound names, CAS-RN, or SMILES, and also offers calculated descriptors using only the SMILES of a compound. To complement this data, we obtained $\log K_{ow}$ values, which reflect hydrophobicity, from the publicly-available Environmental Protection Agency-Estimation Programs Interface (EPA-ESPI) Suite program.[49] This comprehensive set of molecular descriptors enabled systematic characterization of PPCPs.

2.2 Clustering PPCPs based on relative abundance measures obtained from non-target screening

To classify PPCPs according to similarities in their removal patterns along the two treatment trains, we applied two distinct clustering approaches. The first approach was to group PPCPs as a function of which specific treatment process (*e.g.*, activated sludge, BAC, GAC, chlorination) achieved the greatest removal efficiency relative to the PPCP concentration measured in the influent. The second method was to cluster the PPCPs based on the removal pattern across each process relative to the influent to that process (*i.e.*, did the relative abundance increase or decrease relative to the previous treatment step?). These clustering methods are subsequently denoted as 'C1', and 'C2', respectively. Clustering analysis was applied to each individual treatment facility dataset. For preprocessing, PPCPs with missing relative abundance values across all treatment processes were removed and the common PPCPs shared among all events were extracted.

2.2.1 C1: clustering based on the most efficient individual process relative to treatment train influent. The Unit Removal Efficiency (URE) for each PPCP was calculated using eqn (1):

$$\mathrm{URE} = \frac{P_{\mathrm{in}} - P_{\mathrm{out}}}{P_{\mathrm{initial}}} \times 100 \quad (1)$$

where P_{in} in represents the input peak area of a PPCP for treatment process x, P_{out} the output peak area of a PPCP for treatment process x, and $P_{initial}$ is the input peak area of a PPCP in the wastewater influent at each facility. To obtain the representative URE across the four sampling events, for each treatment process, the average of the efficiencies from all events were calculated, where calculations comparing two subsequent below detection measurements were excluded. For each PPCP, the treatment process achieving the highest average initial removal efficiency was selected, and PPCPs sharing this same treatment process were clustered together.

2.2.2 C2: clustering based on the removal pattern of PPCP in a given process relative to the immediately upstream process. The removal efficiency (RE) of each treatment process for each PPCP was calculated following eqn (2):

$$\mathrm{RE} = \frac{P_{\mathrm{in}} - P_{\mathrm{out}}}{P_{\mathrm{in}}} \times 100 \quad (2)$$

where P_{in} and P_{out} are the input and the output peak area of a compound of a treatment process, respectively. The average removal efficiency was determined across the four sampling trips for each treatment process. The average for each treatment process was then compared to that of the previous treatment process and transformed to one of four categorical variables: increase, decrease, same, or below detection (B.D.), resulting in a sequence of categorical variables for each PPCP, representing the overall removal pattern. Then, *K*-modes clustering[50] which is specifically designed for categorical data, was applied to group PPCPs bearing similar removal patterns into clusters, where the K was set as 3.

2.3 Classification of PPCPs using Abraham descriptors and $\log K_{ow}$

PPCPs defined by C1 and C2 were further classified sing the Abraham descriptors and $\log K_{ow}$ values as inputs. Of particular interest was whether the PPCPs in the same cluster shared similar chemical characteristics. For this purpose, three machine learning-based algorithms were applied: support vector machine (SVM),[51] random forest (RF),[52] and logistic regression (LR),[53] implemented based on 'Scikit-learn' package[54] with the default hyperparameter settings. SVM is a supervised learning algorithm that identifies a hyperplane to create a decision boundary classifying the data points to each class by maximizing the margin between the classes. RF constructs multiple tree-structured classification models based on the set of discrete rules from the training dataset and aggregates the output from multiple trees to derive a final prediction output. LR estimates the probability for the given class using a sigmoid function to the output of linear regression function. We selected these three models due to their strong classification performance compared to methods like Naive Bayes and Decision Trees.[55]

2.4 Validation of the ML-based framework to predict the removal of PPCPs movement

To evaluate the proposed ML-based computational framework, several validation experiments were performed. First, we compared the clustering results obtained from C1 and C2 approaches between the two treatment facilities. Then, 5-fold cross validation was performed on the classification model to test the average accuracy, which was used as a metric to estimate whether the PPCPs can be predicted to one of the defined clusters based on the chemical properties. The PPCPs were divided into five parts ("folds"), with each part containing an equal number of PPCPs. In each iteration, four out of the five parts were used for training the ML-based classification models and the remaining part was solely used for testing. This process was repeated five times, with each fold serving as the test set exactly once. This evaluation followed standard cross validation approach[55–57] and served to demonstrate whether the model can accurately predict the remove of new PPCPs that were not included in the training. Additionally, we performed *K*-mean clustering to group PPCPs based on the Abraham descriptor with $\log K_{ow}$ values to obtain the same number of clusters from C1 and C2. Based on the latter, we checked whether there was agreement between the PPCP clustering results using chemical properties and the clustering results based on the relative abundance using our proposed methods to determine whether the PPCPs in the same cluster share similar chemical properties. The overlap between PPCPs and the number/percentage of PPCPs in clusters defined by chemical properties *versus* those defined by remove efficiencies/patterns in the C1/C2 clustering approaches were compared.

2.5 Statistical analysis of the distinct chemical properties across PPCP clusters

Statistical testing was performed to assess whether the identified PPCP clusters showed distinct distributions of chemical properties across clusters, and similar properties within clusters. Using the Abraham descriptors and $\log K_{ow}$ values of each PPCP, Kruskal–Wallis *H*-test was performed by the basic statistic package in R (v4.1.2). Statistical significance was set at p-value <0.05. A user manual for running the proposed framework is provided in ESI† S2.

3. Results and discussion

3.1 Occurrence and removal of PPCPs across the two treatment trains

Analysis of 149 PPCPs across two distinct treatment trains revealed a complex landscape of contaminant behavior and removal efficiencies, providing a rich dataset for developing and validating the PPCP removal prediction approach. Pharmaceuticals dominated the detected compounds (88.6%), with analgesics, antibiotics, and antidepressants showing the highest prevalence (Fig. 2b). The diverse array of compounds, including various therapeutic classes, pharmaceutical metabolites (4.67%), and personal care products (2%), presented a wide range of physicochemical properties, which supported development of robust predictive models.

The two treatment trains demonstrated varying levels of PPCP removal, with the potable reuse system achieving approximately 20% elimination of the screened compounds. The non-potable reuse system exhibited a 10% reduction. This is consistent with expectation of the advanced treatment processes employed in the potable reuse system, such as ozonation and UV, achieving greater overall removal.[58,59]

General patterns of PPCPs removal were consistent with expectation based on molecular structure. For instance, tramadol, is a complex organic compound that includes a tertiary amine and multiple aromatic rings and was found to be particularly recalcitrant. In contrast, chemicals like caffeine and acetaminophen were effectively eliminated, likely through sorption to solids. These observations underscore the potential of molecular descriptors, such as Abraham descriptors and log

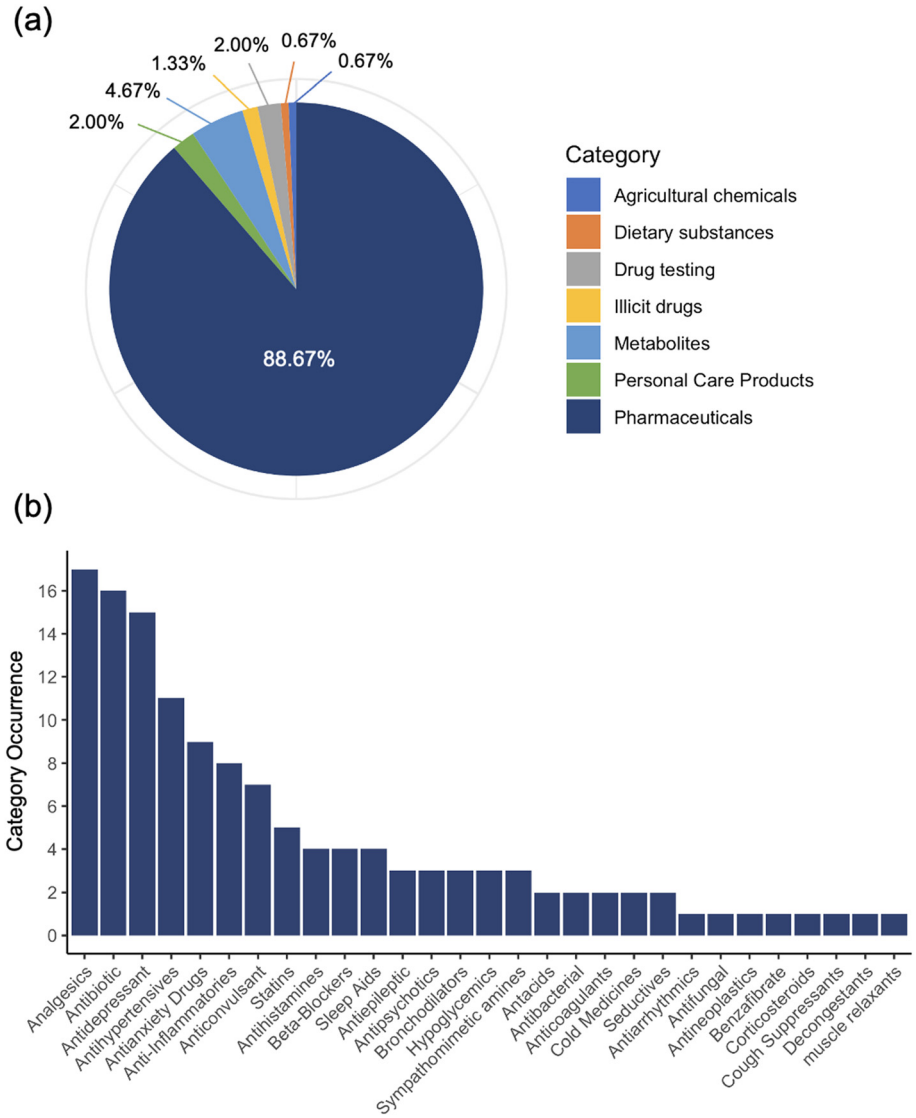


Fig. 2 (a) Distribution of PPCPs across the two wastewater and water reuse treatment trains (b) frequency of occurrence of various categories in the pharmaceutical group across all samples (n = 84) collected from the various stages of treatment for the two treatment trains and across four sampling events.

K_{ow} values, to capture nuanced relationships between chemical properties and efficacy of treatment.

The varying removal efficiencies observed across different treatment stages and between the two systems illuminated complex interplays between PPCP chemical properties and process-specific removal mechanisms. Hydrophobic compounds, exemplified by the anti-epileptic drug carbamazepine, generally persisted through initial treatment stages, but were successfully removed by ozone. Carbamazepine contains an electron-donating amine, which is known to be susceptible to ozonation, illustrating an expected linkage between molecular properties (*i.e.*, containing an electron-donating subgroup) and its susceptibility to treatment (*i.e.*, an electron-attracting oxidative process). Such results are consistent with prior studies that demonstrated the effectiveness of advanced oxidation processes, like ozonation and UV light, in eliminating hydrophobic and other electron-rich PPCPs.[60]

The heterogeneity in PPCP composition and removal patterns observed in this study not only illustrate the challenges of removing them *via* a unified treatment approach, but also highlights the potential for data-driven, predictive models to revolutionize treatment strategy optimization. By capturing nuanced relationships between molecular properties and treatment efficacies, ML models present a promising approach to enhance the ability to predict the fate of PPCPs across various treatment scenarios.

3.2 Quantifying PPCP removal patterns across treatment stages

The C1 clustering approach revealed the relative contribution of each treatment stage to PPCP removal (Fig. 3). Oxidative processes formed the largest clusters, with ozonation being the most efficient removal method for 55.3% of the studied

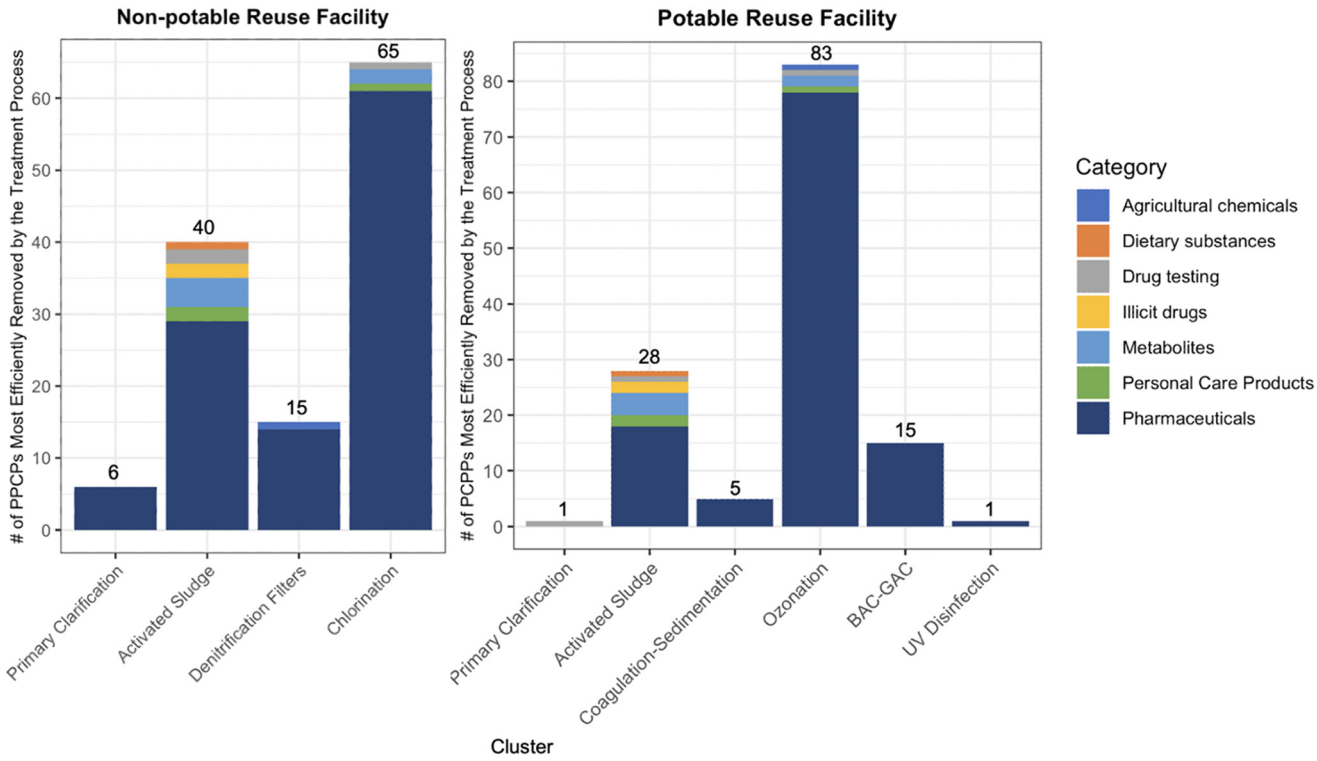


Fig. 3 The number of PPCPs in each cluster based on the C1 clustering method, *i.e.*, according to the most efficient individual treatment process contributing to removal of each PPCP across the treatment train.

PPCPs in the potable reuse system. Chlorination was the primary removal mechanism for 43.3% of PPCPs in the non-potable system. This aligns with previous research highlighting the effectiveness of oxidative processes in degrading a wide range of organic contaminants.[61]

The activated sludge process also played a substantial role, being the most efficient removal stage for 18.7% and 26.6% of PPCPs in potable and non-potable systems respectively. This underscores the importance of biological processes in PPCP degradation,[62] particularly for compounds susceptible to biodegradation. Physical processes, like primary clarification, formed smaller clusters, being the primary removal mechanism for only 4% of PPCPs in the non-potable system and 2% in the potable system. However, these processes still contributed to overall removal, with an average removal efficiency of 15% across all PPCPs. The PPCP lists for each cluster are detailed in Table S1† and the distribution of the removal efficiencies across treatment processes for each cluster are shown in Fig. S1.†

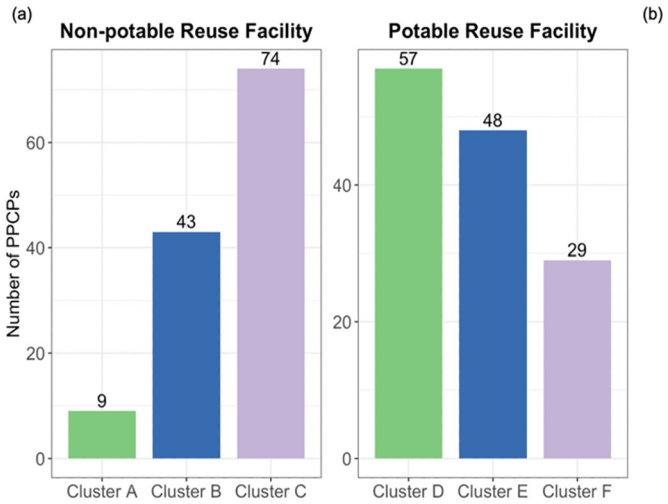
The C2 clustering approach (Fig. 4, Table S2†) provided insight into the cumulative effects of each treatment stage. This analysis revealed that 68% of PPCPs experienced over 90% removal in the latter stages of treatment, particularly during oxidative processes. However, 22% of compounds showed substantial removal (>50%) in earlier stages, highlighting the importance of multi-barrier approaches in wastewater and reuse treatment trains.

It's important to note that removal efficiencies varied considerably among different PPCPs, even within the same treatment stage. For instance, while ozonation showed high removal efficiency (>90%) for 62% of PPCPs, it was less effective (<30% removal) for 18% of compounds. This variability underscores how inherent differences in PPCP physicochemical properties dictate the need for distinct treatment strategies.

These findings provide valuable insights into the relative contributions of different treatment stages to PPCP removal in wastewater and water reuse treatment trains. They highlight the importance of oxidative processes, such as ozone, while also demonstrating the significant role of biological treatment. Furthermore, they emphasize the value of a multi-barrier approach in achieving comprehensive PPCP removal, with each stage contributing to the overall reduction of PPCPs.

3.3 Predicting the removal of PPCPs through water reuse treatment trains using the ML-based classifiers

To predict PPCP removal patterns, we implemented ML-based classification models using Abraham descriptors and $\log K_{ow}$ values as inputs to independently classify each PPCP according to the clusters defined by the C1 and C2 approaches. We employed a five-fold cross-validation method using datasets from each treatment train, measuring classification accuracy on the test dataset to evaluate the models' ability to predict removal patterns for new, unseen PPCPs.



(b)

Non-potable Reuse Facility

Treatment unit	Activated Sludge	Denitrification Filters	Chlorination
Cluster A	Increase	B.D.	B.D.
Cluster B	Increase	Decrease	Increase
Cluster C	Decrease	Increase	Increase

Potable Reuse Facility

Treatment unit	Activated Sludge	Coagulation-Sedimentation	Ozonation	BAF-GAC	UV Disinfection
Cluster D	Increase	Decrease	Increase	Decrease	Decrease
Cluster E	Decrease	Increase	Increase	Increase	Decrease
Cluster F	Decrease	Increase	Increase	B.D.	B.D.

Fig. 4 (a) The number of PPCPs in each cluster based on the C2 clustering approach, *i.e.*, removal pattern across the water reuse trains. Three clustering patterns were observed at each water reuse facility: 'cluster A', 'cluster B', and 'cluster C' for non-potable reuse facility and 'cluster D', 'cluster E', and 'cluster F' for potable reuse facility, where the removal patterns refer to changes in peak area relative abundance. (b) C2 PPCP removal patterns observed for the nonpotable and potable reuse facilities. B.D. – the compound was below detection by the time it reached the corresponding treatment stage. An "increase" and "decrease" were defined to represent changes in the representative removal efficiency of each treatment process, indicating an increase or decrease relative to the efficiency of the previous treatment stage, respectively.

For the prediction of PPCPs to clusters based on the C1 clustering approach (most efficient individual treatment process), random forest (RF) was found to achieve the highest average classification accuracies of 0.539 and 0.652 for the non-potable and potable reuse facilities, respectively (Fig. 5a). Support vector machine (SVM) showed similar performance with average accuracies of 0.522 and 0.652, while logistic regression (LR) yielded accuracies of 0.504 and 0.652 for the respective facilities. These results suggest that Abraham descriptors and $\log K_{ow}$ values capture significant information about a PPCP's susceptibility to specific treatment processes.

In classifying PPCPs according to the C2 clustering approach (removal pattern across all processes), SVM demonstrated the highest average accuracy of 0.597 for the non-potable facility, while achieving 0.425 for the potable facility (Fig. 5b). The second-best performances were observed with LR (0.563) for the non-potable facility and RF (0.424) for the potable facility. This variation in model performance between C1 and C2 classifications highlights complex relationships between molecular properties and overall removal patterns across multiple treatment stages. Additionally, using the random forest classifier, we measured the relative feature importance scores to identify which Abraham descriptors contributed most to predicting PPCP in each cluster for both C1 and C2. The results showed that, in most cases, '$\log K_{ow}$' and 'V' were the two most important features for the prediction task. The exception was

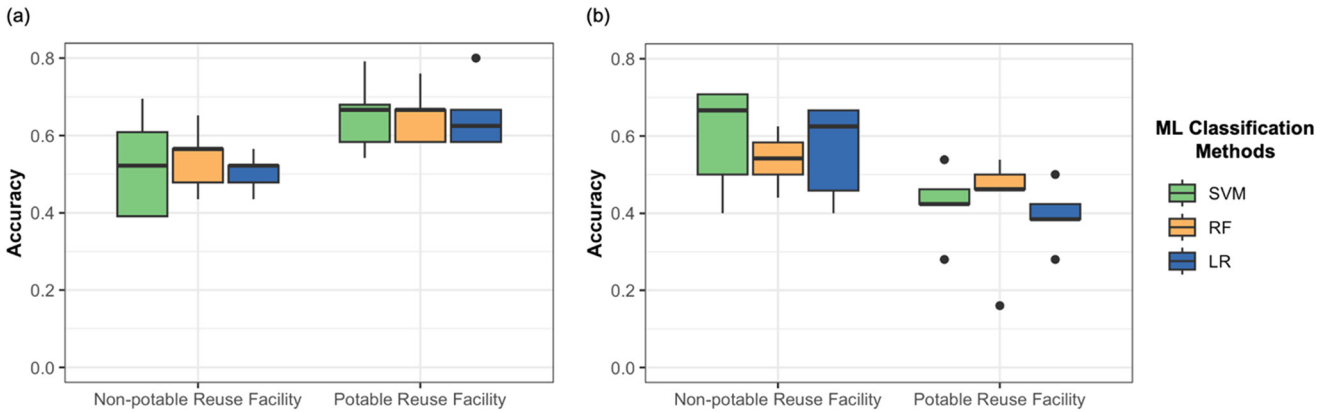


Fig. 5 Classification performance results for predicting PPCPs to the clusters defined based on (a) C1 clustering and (b) C2 clustering approach for each treatment train, performing 5-fold cross validation. 5-Fold cross validation involves dividing the dataset into five subsets, training the model on four of them, and evaluating its performance on the fifth in a cyclic fashion, repeating this process five times to obtain a robust performance estimate. Machine learning-based (ML) classification methods: SVM – support vector machine; RF – random forest; LR – logistic regression. Each boxplot shows the distribution of accuracies for the ML classifiers in 5-fold cross validation of each cluster prediction, where the dot denotes the accuracies detected as the outlier by IQR rule.

for predicting PPCPs into C2-based clusters in the potable reuse facility, where 'E' and '$\log K_{ow}$' were the most significant features.

The approach developed here represents a significant advance in the application of ML techniques to PPCP fate prediction in water reuse systems. While a few studies have used ML for specific aspects of PPCP treatment, such as metal–organic frameworks removal capacity[63] or photocatalytic degradation,[64] our approach is the first to comprehensively characterize PPCP removal across various water reuse treatment processes using molecular descriptors.

The accuracies achieved by our ML models ranged from 42.5% to 65.2% and depended on the facility and clustering approach. While accuracy was moderate, it was not unexpected considering the complex nature of PPCP removal processes and the limited previous applications of ML in this domain. Overall, these results demonstrate the potential of the ML framework to predict PPCP removal patterns based solely on Abraham descriptors and $\log K_{ow}$ values, which can be improved upon in the future. The findings further support the hypothesis that PPCP physicochemical properties can predict PPCP response to various treatment processes.

The variation in model performance between C1 and C2 classifications and between facilities suggests that different ML algorithms may be more suitable for specific aspects of PPCP removal prediction. This underscores the value of our multi-model approach in capturing diverse aspects of PPCP behavior in water reuse systems, aligning with our aim to identify underlying associations between physicochemical properties and removal patterns.

Furthermore, the ability of our models to achieve good accuracy using only Abraham descriptors and $\log K_{ow}$ as inputs is particularly noteworthy. It suggests that molecular properties, specifically, and physicochemical properties generally, are indeed relevant predictors of PPCP fate in water reuse treatment processes, validating our approach of using these descriptors to classify PPCPs according to their defined clusters.

The findings of this study not only advance the application of ML in PPCP removal prediction, but also provide a foundation for future refinements and expansions of this approach. By demonstrating the feasibility of using ML to characterize PPCP removal in wastewater and water reuse facilities based on molecular descriptors and other physicochemical properties, our study opens new avenues for optimizing treatment processes and assessing the fate of emerging contaminants.

3.4 Validation of distinct physicochemical properties across PPCP clusters

In addition to the 5-fold cross validation in section 3.3, we investigated whether the PPCPs in each cluster shared similar physicochemical properties and distinct distributions compared to other clusters. Abraham descriptor and $\log K_{ow}$ values are compared by cluster in Fig. S2 and S3.† It was found that, Abraham descriptors A and E were significantly different across the C1 clusters for both treatment facilities (Kruskal–Wallis, p-value <0.05). However, no significant difference in these values was found across the C2 clusters (Fig. S4†).

We further performed K-mean clustering to group PPCPs based on the Abraham descriptor and $\log K_{ow}$ values and checked whether there was agreement between the PPCP clustering results using physicochemical properties *versus* removal efficiencies/patterns defined by the C1/C2 clustering approaches. It was found that 61 and 71 PPCPs overlapped in

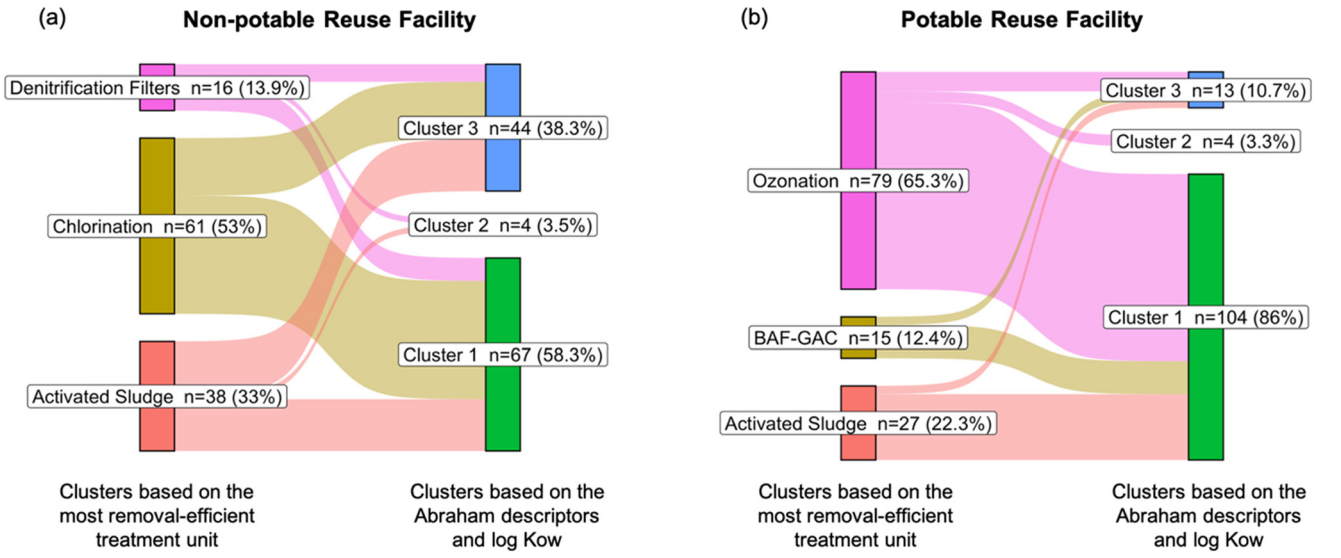


Fig. 6 Number of PPCPs overlapping among the clusters based on C1 clustering, *i.e.*, the treatment process that achieved the most efficient removal, and based on the physicochemical properties for (a) non-potable and (b) potable reuse facility.

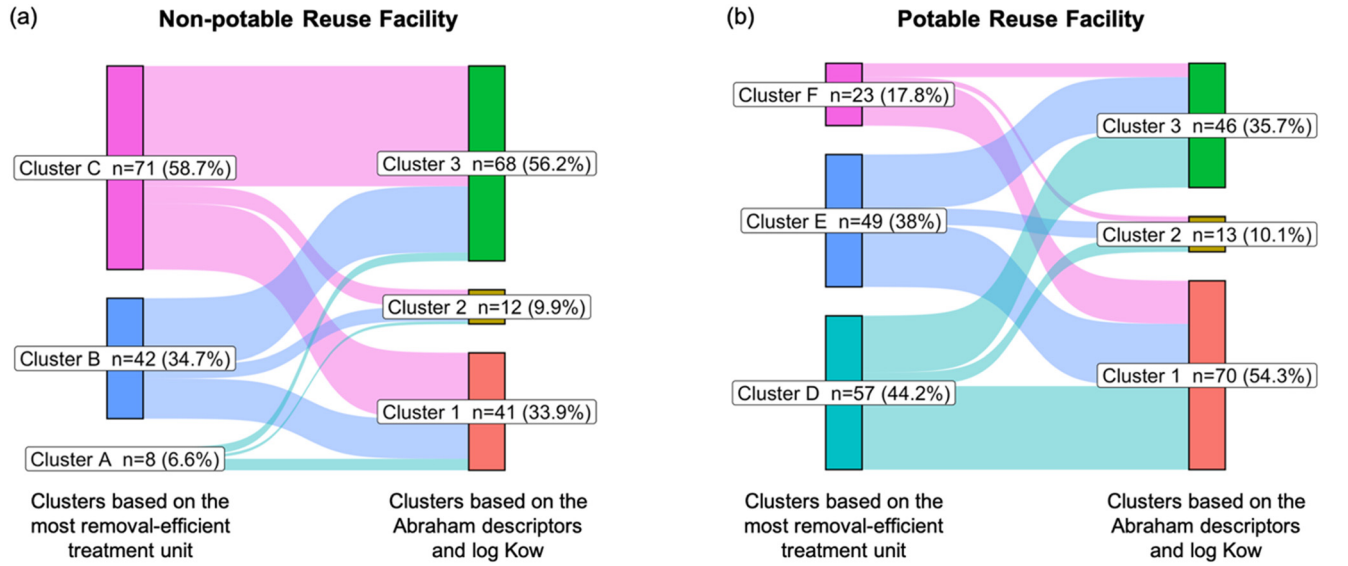


Fig. 7 Number of PPCPs overlapping among the C2 clusters, based on the removal pattern across the facility, and clusters based on the physicochemical properties for (a) non-potable and (b) potable reuse facility.

their groupings based on C1 clustering at the non-potable and potable treatment facilities, respectively (Fig. 6). For the C2 clustering approach, 56 and 54 PPCPs overlapped in their classification based on the Abraham descriptor and $\log K_{ow}$ values (Fig. 7). The full list of overlapping PPCPs is reported in Tables S4 and S5.†

Among the C1 clustering overlaps, PPCPs were not dominated by a single category, but represented a diverse range of pharmaceuticals. However, analgesics, antidepressants, antianxiety drugs, and antihypertensives were more prevalent among the overlapping clusters (Table S6†). These compounds predominantly clustered under oxidative processes, *i.e.*, ozonation for potable reuse systems and chlorination for non-potable reuse systems. Upon further analysis of Abraham descriptors represented by these clusters (Fig. S2†), we found that PPCPs in these clusters exhibited higher A, B (hydrogen bond basicity), and $\log K_{ow}$ values compared to other clusters. This indicates a trend of higher hydrophobicity, which supports more effective removal through advanced oxidation processes such as ozonation and chlorination.[65]

Among the C2 clustering, we observed a comparable dominance of the PPCP categories in the overlapping clusters. Most of these PPCPs were found in cluster C for non-potable systems and clusters D and E for potable systems. In terms of treatment trends, chlorination (for non-potable systems) and ozonation (for potable systems) again achieved greater percent removal compared to earlier treatment stages. Further analysis of Abraham descriptors for these clusters (Fig. S3†) revealed similar trends in chemical properties, with PPCPs in these clusters having higher A, B, and $\log K_{ow}$ values. This further suggests that higher hydrophobicity and these specific chemical properties contribute to the enhanced removal efficiency observed in both ozonation and chlorination processes.[65]

4. Future directions

This study demonstrates a promising avenue to predict the removal of emerging, previously uncharacterized, PPCPs through various candidate process treatment processes, based on their physicochemical properties. This could be a valuable approach towards treatment train design and operation for maximal removal. While promising, further refinement and testing of the approach developed herein would be beneficial. To move towards developing more accurate predictive models for PPCP removal, expanding available databases summarizing key PPCP physicochemical parameters would be of value, including biodegradation kinetics, reaction rate constants, chemical structures, and sorption coefficients. Incorporating localized temporal variation in environmental conditions such as temperature, pH, or rainfall could also help account for variability in concentration trends. Changes in such factors are known to influence contaminant degradation and transport mechanisms.[66] Furthermore, operational parameters such as differing dissolved oxygen concentrations, solids retention times, and flowrates used across different facilities could explain some of the variance observed between facilities. With these data resources in place, advanced deep learning algorithms capable of capturing nonlinear relationships, such as multi-layer artificial neural networks can be implemented to relate PPCP properties and influent concentrations to effluent concentrations and removal efficiencies across treatment steps.

Emerging generative modeling techniques could be one avenue for overcoming key data limitations. Following model development, further validation using data from additional treatment trains not included in the training data set would be of value. Once validated, user-friendly tools could be

developed for consultants, plant operators and regulators. By inputting PPCP properties and operating conditions, the models could efficiently predict expected removal and guide the design and operation of corresponding treatment trains. Applications could include risk assessment of new chemicals and *in silico* screening prior to market entry. Overall, leveraging ML on expanded PPCP data has potential to enable predictive approaches that bolster wastewater and water reuse treatment and management strategies.

5. Limitations of sampling approach and considerations

This study provides valuable insights into the observed differences between influent and effluent compositions across and advanced water treatment train. Sampling was conducted over four events in order to capture general removal trends over time. However, ideally, sampling could have been more precisely timed to account for hydraulic retention time (HRT) and to attempt to follow the same parcel of water through each unit treatment process. Considering this limitation, the term 'removals' used throughout the manuscript should be interpreted as indicative of typical differences in influent *versus* effluent during stable operation, rather than as definitive removal efficiencies. Future studies could consider timing sampling in a manner that takes into account HRT as a means to account for temporal variation in treatment dynamics and enhance the precision of removal estimates.

6. Conclusion

In this study, 149 PPCPs were screened through wastewater treatment and subsequent potable and non-potable water reuse treatment trains using non-targeted UPLC-MS/MS analysis to evaluate efficacy of various physical, biological, oxidative, and sorptive treatment processes for their removal. PPCPs were clustered based on the relative abundances measured through each treatment step using two approaches: C1 grouped PPCPs based on their most efficient individual treatment process, while C2 clustered PPCPs according to their removal pattern across the treatment train. ML-based classification algorithms including SVM, RF, and LR were applied to relate PPCP physicochemical descriptors to their cluster assignments. The results suggested that PPCPs within each cluster generally share similar physicochemical properties, as reflected by similarities among Abraham descriptors E, S, A, B, and V. Further, each cluster has distinct characteristics from one another. The C1 clustering provides insight into the most suitable treatment technology for specific PPCPs. Meanwhile, the C2 clustering elucidates general trends of PPCP persistence and removal in reuse systems.

Here, a novel framework for predicting PPCP removal by various treatment processes was developed combining supervised and unsupervised ML and informed by specific physicochemical properties of each PPCP. This study demonstrated the ability of ML techniques; RF, SVM and LR, to systematically characterize and classify PPCP removal, using extensive PPCP screening data sets collected through two wastewater treatment plants followed by distinct water reuse treatment trains. Looking forward, considering additional molecular descriptors, and utilizing more advanced ML techniques and drawing from a broader array of data sets can help to further develop this framework into a practical, accurate tool for consultants, operators and regulators. The framework developed here could be of particular value for informing the design of water reuse treatment trains to meet ever growing demands for removal of a broader array of PPCPs, including emerging contaminants of concern. The approach could help to complement and amplify the value of costly direct testing and monitoring of PPCPs in wastewater and reuse treatment trains.

Data availability

The complete framework of the code used in this study is provided in ESI† S2. All major data related to this study are reported in the supplementary tables and figures. Additionally, the complete dataset can be accessed at **https://github.com/joungmin-choi/ML-PPCP**.

Author contributions

Development and implementation of the data analysis approach were performed by JMC and VM. Sampling strategy and collection was performed by IK, who also assisted with preliminary analysis of the data. CC performed the PPCP analysis under supervision of KX. The first draft of the manuscript was written by JMC and VM. AP, LZ, KX, PJV, MFB, CB and CB contributed to conceptualization of the study, supervision, review and editing. All authors read and approved the final manuscript.

Conflicts of interest

Other than funding provided by the employer of co-author Charles Bott, 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.

Acknowledgements

This work was supported in part by the U.S. National Science Foundation Awards 2004751 and 2125798, U.S. Centers for Disease Control and Prevention contract 75D30118C02904, US Bureau of Reclamation grant R21AC10162-01, EPA Grant R840619, funding from the Hampton Roads Sanitation District, and a Pilot Program to Enhance NIH Funding within the COE at Virginia Tech. The authors would like to thank Drs. Chao Shang and Sheldon Hilaire for the PPCP screening analysis and the staff at the water utilities for assisting with sample collection.

References

1 P. Loganathan, S. Vigneswaran, J. Kandasamy, A. K. Cuprys, Z. Maletskyi and H. Ratnaweera, Treatment Trends and Combined Methods in Removing Pharmaceuticals and Personal Care Products from Wastewater—A Review, *Membranes*, 2023, **13**, 158.

2 Y. Yang, Y. S. Ok, K.-H. Kim, E. E. Kwon and Y. F. Tsang, Occurrences and removal of pharmaceuticals and personal care products (PPCPs) in drinking water and water/sewage treatment plants: A review, *Sci. Total Environ.*, 2017, **596–597**, 303–320.

3 T. Aus Der Beek, F. Weber, A. Bergmann, S. Hickmann, I. Ebert, A. Hein and A. Küster, Pharmaceuticals in the environment—Global occurrences and perspectives, *Environ. Toxicol. Chem.*, 2016, **35**, 823–835.

4 B. J. Richardson, P. K. S. Lam and M. Martin, Emerging chemicals of concern: Pharmaceuticals and personal care products (PPCPs) in Asia, with particular reference to Southern China, *Mar. Pollut. Bull.*, 2005, **50**, 913–920.

5 J. Wilkinson, P. S. Hooda, J. Barker, S. Barton and J. Swinden, Occurrence, fate and transformation of emerging contaminants in water: An overarching review of the field, *Environ. Pollut.*, 2017, **231**, 954–970.

6 Z. Tousova, P. Oswald, J. Slobodnik, L. Blaha, M. Muz, M. Hu, W. Brack, M. Krauss, C. Di Paolo, Z. Tarcai, T.-B. Seiler, H. Hollert, S. Koprivica, M. Ahel, J. E. Schollée, J. Hollender, M. J.-F. Suter, A. O. Hidasi, K. Schirmer, M. Sonavane, S. Ait-Aissa, N. Creusot, F. Brion, J. Froment, A. C. Almeida, K. Thomas, K. E. Tollefsen, S. Tufi, X. Ouyang, P. Leonards, M. Lamoree, V. O. Torrens, A. Kolkman, M. Schriks, P. Spirhanzlova, A. Tindall and T. Schulze, European demonstration program on the effect-based and chemical identification and monitoring of organic pollutants in European surface waters, *Sci. Total Environ.*, 2017, **601–602**, 1849–1868.

7 P. Chaturvedi, P. Shukla, B. S. Giri, P. Chowdhary, R. Chandra, P. Gupta and A. Pandey, Prevalence and hazardous impact of pharmaceutical and personal care products and antibiotics in environment: A review on emerging contaminants, *Environ. Res.*, 2021, **194**, 110664.

8 L. Yang, Y. Zhou, B. Shi, J. Meng, B. He, H. Yang, S. J. Yoon, T. Kim, B.-O. Kwon, J. S. Khim and T. Wang, Anthropogenic impacts on the contamination of pharmaceuticals and personal care products (PPCPs) in the coastal environments of the Yellow and Bohai seas, *Environ. Int.*, 2020, **135**, 105306.

9 A. S. Adeleye, J. Xue, Y. Zhao, A. A. Taylor, J. E. Zenobio, Y. Sun, Z. Han, O. A. Salawu and Y. Zhu, Abundance, fate, and effects of pharmaceuticals and personal care products in aquatic environments, *J. Hazard. Mater.*, 2022, **424**, 127284.

10 M. Zhang, J. Shen, Y. Zhong, T. Ding, P. D. Dissanayake, Y. Yang, Y. F. Tsang and Y. S. Ok, Sorption of pharmaceuticals and personal care products (PPCPs) from water and wastewater by carbonaceous materials: A review, *Crit. Rev. Environ. Sci. Technol.*, 2022, **52**, 727–766.

11 L. Cizmas, V. K. Sharma, C. M. Gray and T. J. McDonald, Pharmaceuticals and personal care products in waters: occurrence, toxicity, and risk, *Environ. Chem. Lett.*, 2015, **13**, 381–394.

12 A. Pal, K. Y.-H. Gin, A. Y.-C. Lin and M. Reinhard, Impacts of emerging organic contaminants on freshwater resources: Review of recent occurrences, sources, fate and effects, *Sci. Total Environ.*, 2010, **408**, 6062–6069.

13 U. Anand, B. Adelodun, C. Cabreros, P. Kumar, S. Suresh, A. Dey, F. Ballesteros and E. Bontempi, Occurrence, transformation, bioaccumulation, risk and analysis of pharmaceutical and personal care products from wastewater: a review, *Environ. Chem. Lett.*, 2022, **20**, 3883–3904.

14 C. Yan, J. Jin, J. Wang, F. Zhang, Y. Tian, C. Liu, F. Zhang, L. Cao, Y. Zhou and Q. Han, Metal–organic frameworks (MOFs) for the efficient removal of contaminants from water: Underlying mechanisms, recent advances, challenges, and future prospects, *Coord. Chem. Rev.*, 2022, **468**, 214595.

15 J. Ryu, J. Oh, S. A. Snyder and Y. Yoon, Determination of micropollutants in combined sewer overflows and their removal in a wastewater treatment plant (Seoul, South Korea), *Environ. Monit. Assess.*, 2014, **186**, 3239–3251.

16 J. Wang and S. Wang, Removal of pharmaceuticals and personal care products (PPCPs) from wastewater: A review, *J. Environ. Manage.*, 2016, **182**, 620–640.

17 N. Bolong, A. F. Ismail, M. R. Salim and T. Matsuura, A review of the effects of emerging contaminants in wastewater and options for their removal, *Desalination*, 2009, **239**, 229–246.

18 M. Dubey, B. P. Vellanki and A. A. Kazmi, Emerging contaminants in conventional and advanced biological nutrient removal based wastewater treatment plants in India- insights into the removal processes, *Sci. Total Environ.*, 2023, **894**, 165094.

19 H. Wei, M. Tang and X. Xu, Mechanism of uptake, accumulation, transport, metabolism and phytotoxic effects of pharmaceuticals and personal care products within plants: A review, *Sci. Total Environ.*, 2023, **892**, 164413.

20 H. Wei, M. Tang and X. Xu, Mechanism and influence factors of plant uptake, accumulation, transport, metabolism pathways of pharmaceuticals and personal care products and their phytotoxicity: A review, *Sci. Total Environ.*, 2023, 164413.

21 J. Cheng, H. Du, M.-S. Zhou, Y. Ji, Y.-Q. Xie, H.-B. Huang, S.-H. Zhang, F. Li, L. Xiang, Q.-Y. Cai, Y.-W. Li, H. Li, M. Li, H.-M. Zhao and C.-H. Mo, Substrate-enzyme interactions and catalytic mechanism in a novel family VI esterase with dibutyl phthalate-hydrolyzing activity, *Environ. Int.*, 2023, **178**, 108054.

22 S. Jamil, P. Loganathan, S. J. Khan, J. A. McDonald, J. Kandasamy and S. Vigneswaran, Enhanced nanofiltration rejection of inorganic and organic compounds from a wastewater-reclamation plant's micro-filtered water using adsorption pre-treatment, *Sep. Purif. Technol.*, 2021, **260**, 118207.

23 N. K. Khanzada, M. U. Farid, J. A. Kharraz, J. Choi, C. Y. Tang, L. D. Nghiem, A. Jang and A. K. An, Removal of organic micropollutants using advanced membrane-based water and wastewater treatment: A review, *J. Membr. Sci.*, 2020, **598**, 117672.

24 Y. Zhao, S. Lin, J.-W. Choi, J. K. Bediako, M.-H. Song, J.-A. Kim, C.-W. Cho and Y.-S. Yun, Prediction of adsorption properties for ionic and neutral pharmaceuticals and pharmaceutical intermediates on activated charcoal from aqueous solution via LFER model, *Chem. Eng. J.*, 2019, **362**, 199–206.

25 H. Liu, K. Wei, Y. Yu and C. Long, Predicting adsorption coefficients of VOCs using polyparameter linear free energy relationship based on the evaluation of dispersive and specific interactions, *Environ. Pollut.*, 2019, **255**, 113224.

26 K. Zhang, S. Zhong and H. Zhang, Predicting Aqueous Adsorption of Organic Compounds onto Biochars, Carbon Nanotubes, Granular Activated Carbons, and Resins with Machine Learning, *Environ. Sci. Technol.*, 2020, **54**, 7008–7018.

27 X. Zhu, M. He, Y. Sun, Z. Xu, Z. Wan, D. Hou, D. S. Alessi and D. C. W. Tsang, Insights into the adsorption of pharmaceuticals and personal care products (PPCPs) on biochar and activated carbon with the aid of machine learning, *J. Hazard. Mater.*, 2022, **423**, 127060.

28 C. F. Poole, The effect of descriptor database selection on the physicochemical characterization and prediction of water-air, octanol-air and octanol-water partition constants using the solvation parameter model, *J. Chromatogr. A*, 2023, **1706**, 464213.

29 C. Hao, X. Zhao and P. Yang, GC-MS and HPLC-MS analysis of bioactive pharmaceuticals and personal-care products in environmental matrices, *TrAC, Trends Anal. Chem.*, 2007, **26**, 569–580.

30 N. Pérez-Lemus, R. López-Serna, S. I. Pérez-Elvira and E. Barrado, Analytical methodologies for the determination of pharmaceuticals and personal care products (PPCPs) in sewage sludge: A critical review, *Anal. Chim. Acta*, 2019, **1083**, 19–40.

31 H. Qin, H. Liu, Y. Liu, S. Di, Y. Bao, Y. Zhai and S. Zhu, Recent advances in sample preparation and chromatographic analysis of pharmaceuticals and personal care products in environment, *TrAC, Trends Anal. Chem.*, 2023, **164**, 117112.

32 H. T. Trinh, P. Adriaens, C. M. Lastoskie and Department of Civil and Environmental Engineering, The University of Michigan, 1351 Beal Avenue, Ann Arbor, Michigan 48109-2125, USA, Fate factors and emission flux estimates for emerging contaminants in surface waters, *AIMS Environ. Sci.*, 2016, **3**, 21–44.

33 C. Lindim, J. Van Gils and I. T. Cousins, A large-scale model for simulating the fate & transport of organic contaminants in river basins, *Chemosphere*, 2016, **144**, 803–810.

34 S. Song, C. Su, Y. Lu, T. Wang, Y. Zhang and S. Liu, Urban and rural transport of semivolatile organic compounds at regional scale: A multimedia model approach, *J. Environ. Sci.*, 2016, **39**, 228–241.

35 C. I. Kosma, D. A. Lambropoulou and T. A. Albanis, Occurrence and removal of PPCPs in municipal and hospital wastewaters in Greece, *J. Hazard. Mater.*, 2010, **179**, 804–817.

36 M. Ashraf, S. Z. Ahammad and S. Chakma, Advancements in the dominion of fate and transport of pharmaceuticals and personal care products in the environment—a bibliometric study, *Environ. Sci. Pollut. Res.*, 2023, **30**, 64313–64341.

37 S. U. Gerbersdorf, C. Cimatoribus, H. Class, K.-H. Engesser, S. Helbich, H. Hollert, C. Lange, M. Kranert, J. Metzger, W. Nowak, T.-B. Seiler, K. Steger, H. Steinmetz and S. Wieprecht, Anthropogenic Trace Compounds (ATCs) in aquatic habitats — Research needs on sources, fate, detection and toxicity to ensure timely elimination strategies and risk management, *Environ. Int.*, 2015, **79**, 85–105.

38 S. Gupta, D. Aga, A. Pruden, L. Zhang and P. Vikesland, Data Analytics for Environmental Science and Engineering Research, *Environ. Sci. Technol.*, 2021, **55**, 10895–10907.

39 X. Liu, D. Lu, A. Zhang, Q. Liu and G. Jiang, Data-Driven Machine Learning in Environmental Pollution: Gains and Problems, *Environ. Sci. Technol.*, 2022, **56**, 2124–2133.

40 X. C. Nguyen, Q. V. Ly, T. T. H. Nguyen, H. T. T. Ngo, Y. Hu and Z. Zhang, Potential application of machine learning for exploring adsorption mechanisms of pharmaceuticals onto biochars, *Chemosphere*, 2022, **287**, 132203.

41 S. Zhong, K. Zhang, M. Bagheri, J. G. Burken, A. Gu, B. Li, X. Ma, B. L. Marrone, Z. J. Ren, J. Schrier, W. Shi, H. Tan, T. Wang, X. Wang, B. M. Wong, X. Xiao, X. Yu, J.-J. Zhu and H. Zhang, Machine Learning: New Ideas and Tools in Environmental Science and Engineering, *Environ. Sci. Technol.*, 2021, acs.est.1c01339.

42 P. Masuodi, F. Bahmanzadegan, A. Hemmati and A. Ghaemi, Evaluating the efficiency of nanofiltration and reverse osmosis membranes for the removal of micro-pollutants using a machine learning approach, *Case Stud. Chem. Environ. Eng.*, 2024, **9**, 100750.

43 S. J. Lim, J. Seo, M. G. Seid, J. Lee, W. W. Ejerssa, D.-H. Lee, E. Jeong, S. H. Chae, Y. Lee, M. Son and S. W. Hong, Clustering micropollutants and estimating rate constants of sorption and biodegradation using machine learning approaches, *npj Clean Water*, 2023, **6**, 1–10.

44 N. D. Viet and A. Jang, Machine learning-based real-time prediction of micropollutant behaviour in forward osmosis membrane (waste)water treatment, *J. Cleaner Prod.*, 2023, **389**, 136023.

45 I. Keenum, J. Calarco, H. Majeed, E. Hager, C. Bott, E. Garner, V. J. Harwood and A. Pruden, To what Extent do Water Reuse Treatments Reduce Antibiotic Resistance Indicators? A Comparison of Two Full-Scale Systems, *Water Res.*, 2024, 121425.

46 D.-H. Yang, M. Murphy and S. Zhang, Highly Sensitive Detection of Pharmaceuticals and Personal Care Products (PPCPs) in Water Using an Agilent 6495 Triple Quadrupole Mass Spectrometer Application Note, *Agil. Appl. Note*.

47 L. F. Angeles and D. S. Aga, Establishing Analytical Performance Criteria for the Global Reconnaissance of Antibiotics and Other Pharmaceutical Residues in the Aquatic

Environment Using Liquid Chromatography-Tandem Mass Spectrometry, *J. Anal. Methods Chem.*, 2018, **2018**, 1–9.

48 N. Ulrich, S. Endo, T. N. Brown, N. Watanabe, G. Bronner, M. H. Abraham and K. U. Goss, *UFZ-LSER database v 3.2. 1 [Internet]*, Helmholtz Cent. Environ. Res., Leipz.

49 O. US EPA, Download EPI Suite™ - Estimation Program Interface v4.11, **https://www.epa.gov/tsca-screening-tools/download-epi-suitetm-estimation-program-interface-v411**, (accessed 26 November 2023).

50 A. Chaturvedi, P. E. Green and J. D. Caroll, K-modes Clustering, *J. Classif.*, 2001, **18**, 35–55.

51 W. S. Noble, What is a support vector machine?, *Nat. Biotechnol.*, 2006, **24**, 1565–1567.

52 Y. Qi, in *Ensemble Machine Learning*, ed. C. Zhang and Y. Ma, Springer New York, New York, NY, 2012, pp. 307–323.

53 M. P. LaValley, Logistic Regression, *Circulation*, 2008, **117**, 2395–2399.

54 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

55 J. Roostaei, S. Colley, R. Mulhern, A. A. May and J. M. Gibson, Predicting the risk of GenX contamination in private well water using a machine-learned Bayesian network model, *J. Hazard. Mater.*, 2021, **411**, 125075.

56 R. Du, Q. Zhang, B. Wang, J. Huang, S. Deng and G. Yu, Quantitative structure-activity relationship models for the reaction rate coefficients between dissolved organic matter and PPCPs, *J. Hazard. Mater.*, 2023, **458**, 131845.

57 M. Yaqub, N. M. Ngoc, S. Park and W. Lee, Predictive modeling of pharmaceutical product removal by a managed aquifer recharge system: Comparison and optimization of models using ensemble learners, *J. Environ. Manage.*, 2022, **324**, 116345.

58 P. R. Rout, T. C. Zhang, P. Bhunia and R. Y. Surampalli, Treatment technologies for emerging contaminants in wastewater treatment plants: A review, *Sci. Total Environ.*, 2021, **753**, 141990.

59 Y. Luo, W. Guo, H. H. Ngo, L. D. Nghiem, F. I. Hai, J. Zhang, S. Liang and X. C. Wang, A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment, *Sci. Total Environ.*, 2014, **473–474**, 619–641.

60 A. Chauhan, D. Sillu and S. Agnihotri, Removal of Pharmaceutical Contaminants in Wastewater Using Nanomaterials: A Comprehensive Review, *Curr. Drug Metab.*, 2019, **20**, 483–505.

61 P. V. Nidheesh, C. Couras, A. V. Karim and H. Nadais, A review of integrated advanced oxidation processes and biological processes for organic pollutant removal, *Chem. Eng. Commun.*, 2022, **209**, 390–432.

62 M. Narayanan, M. El-sheekh, Y. Ma, A. Pugazhendhi, D. Natarajan, G. Kandasamy, R. Raja, R. M. Saravana Kumar, S. Kumarasamy, G. Sathiyan, R. Geetha, B. Paulraj, G. Liu and S. Kandasamy, Current status of microbes involved in the degradation of pharmaceutical and personal care products (PPCPs) pollutants in the aquatic ecosystem, *Environ. Pollut.*, 2022, **300**, 118922.

63 R. Xin, C. Wang, Y. Zhang, R. Peng, R. Li, J. Wang, Y. Mao, X. Zhu, W. Zhu, M. Kim, H. N. Nam and Y. Yamauchi, Efficient Removal of Greenhouse Gases: Machine Learning-Assisted Exploration of Metal–Organic Framework Space, *ACS Nano*, 2024, **18**, 19403–19422.

64 A. Gordanshekan, S. Arabian, A. R. Solaimany Nazar, M. Farhadian and S. Tangestaninejad, A comprehensive comparison of green Bi2WO6/g-C3N4 and Bi2WO6/TiO2 S-scheme heterojunctions for photocatalytic adsorption/degradation of Cefixime: Artificial neural network, degradation pathway, and toxicity estimation, *Chem. Eng. J.*, 2023, **451**, 139067.

65 H. Erdem and M. Erdem, Effect of log Kow on the degradation of pharmaceutically active compounds in a heterogeneous catalytic persulfate activation system, *J. Environ. Chem. Eng.*, 2024, **12**, 111720.

66 X. Yu, Q. Sui, S. Lyu, W. Zhao, D. Wu, G. Yu and D. Barcelo, Rainfall Influences Occurrence of Pharmaceutical and Personal Care Products in Landfill Leachates: Evidence from Seasonal Variations and Extreme Rainfall Episodes, *Environ. Sci. Technol.*, 2021, **55**, 4822–4830.