

## ARTICLE OPEN



# Clustering micropollutants and estimating rate constants of sorption and biodegradation using machine learning approaches

Seung Ji Lim<sup>1,5</sup>, Jangwon Seo<sup>1,5</sup>, Mingizem Gashaw Seid<sup>1</sup>, Jiho Lee<sup>1</sup>, Wondesen Workneh Ejerssa<sup>1,2</sup>, Doo-Hee Lee<sup>3</sup>, Eunhoo Jeong<sup>1</sup>, Sung Ho Chae<sup>1</sup>, Yunho Lee<sup>4</sup>, Moon Son<sup>1,2,6</sup> and Seok Won Hong<sup>1,2,6</sup>

Effluent from wastewater treatment plants is considered an important source of micropollutants (MPs) in aquatic environments. However, monitoring MPs in effluents is often inefficient owing to the variety in their types. Thus, this study derived marker constituents to estimate the behavior of MPs in each cluster using the self-organizing map (SOM), a machine learning-based clustering analysis method. In SOM analysis, the physicochemical properties, functional groups, and the initial biotransformation rules of 29 out of 42 MPs were used to ultimately estimate the degradation rate constants of 13 MPs. Consequently, when the physicochemical properties and functional groups were considered, SOM analysis showed outstanding performance to label MPs with an accuracy value of 0.75 for each aerobic and anoxic condition. Based on the clustering results, 11 MPs were determined to be marker constituents under each aerobic and anoxic condition. Moreover, an estimation method for the rate constants of unlabeled MPs was successfully developed using the identified markers with the random forest classifier. The proposed algorithm could estimate both sorption and biotransformation of MPs regardless of dominant removal mechanisms, whether the MPs were removed by sorption or biotransformation. An accuracy of 0.77 was calculated for estimating rate constants under both aerobic and anoxic conditions, which is remarkably higher than those reported previously. The proposed procedure could be extended further to efficiently monitor MPs in effluents.

*npj Clean Water* (2023)6:69; <https://doi.org/10.1038/s41545-023-00282-6>

## INTRODUCTION

Domestic and industrial chemicals, such as pharmaceuticals, personal care products, steroids, estrogens, pesticides, and surfactants have become essential for modern living. More than 350,000 chemicals have been registered and used worldwide from 2010 to 2019<sup>1</sup>. It was also reported that global medicine consumption in 2020 was estimated to be 4.5 trillion doses<sup>2</sup>. Most or some of these chemicals flow into wastewater treatment plants (WWTPs), but are not completely degraded<sup>3,4</sup>. These undecomposed chemicals can adversely impact aquatic ecosystems, when discharged from the WWTPs<sup>5</sup>. To assess the impact and potential risks of these unwanted chemicals, called micropollutants (MPs), frequent and accurate monitoring of effluent from WWTPs is a crucial requirement. However, periodic monitoring of MPs is expensive and labor-intensive.

Rather than monitoring individual MPs, monitoring a grouping of them is more efficient because it can reduce the number of samples to be monitored by providing the group representative values<sup>6,7</sup>. For example, one study selected caffeine, which is widely available in food, drinks, and pharmaceuticals, as a marker for evaluating the degree of aquatic ecosystem contamination by untreated wastewater<sup>8</sup>. The concentration of caffeine in untreated wastewater is orders of magnitude higher than that in treated water from WWTPs due to the high removal efficiency of caffeine during wastewater treatment in general (>99%). Therefore, rather than detecting individual MPs, solely monitoring the caffeine

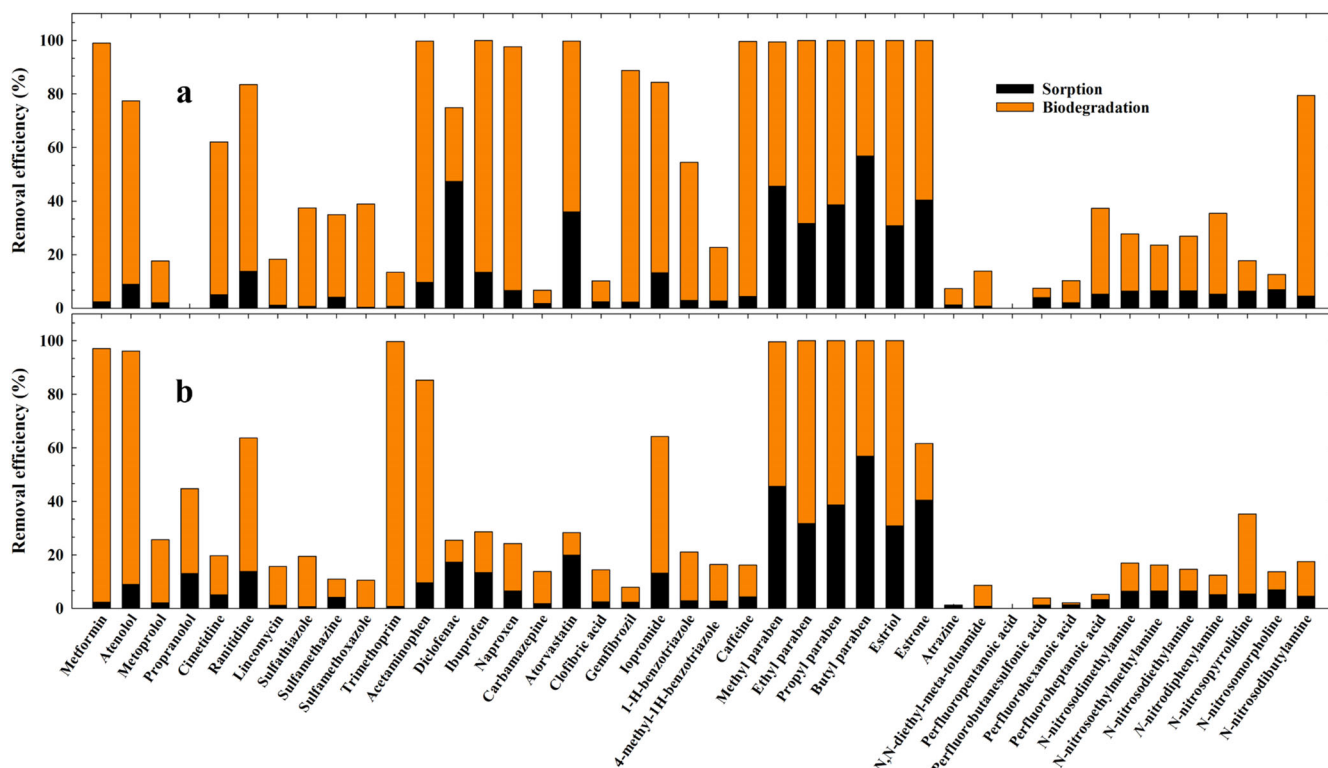
concentration was sufficient for assessing anthropogenic contamination<sup>9</sup>. Similar to the monitoring of MPs in the aquatic ecosystem, their monitoring in WWTPs would be significantly simplified if a representative marker for each group of MPs is selected prior to analyzing the concentration of all MPs.

Several clustering analyses have been attempted in order to identify similarities among MPs and increase the prediction accuracy of their behavior. One example is the dendrogram, which is often employed to generate a graphical representation exhibiting the trends of biodegradation rate constants associated with solid retention time<sup>10,11</sup>. Recently, the clustering of MPs using initial biotransformation rules was also introduced using the Eawag pathway prediction system (Eawag-PPS)<sup>12,13</sup>. Although these clustering methods are suitable (i.e., wide range of applications or explainable clustering), they contain issues that need to be addressed. For instance, the dendrogram provides insufficient information for interpreting the clustering results since it represents clustering results on the one-dimensional graph. While the clustering with the biotransformation rule is more explainable, it often shows insufficient prediction accuracy because it lacks detailed chemical characteristics such as functional groups<sup>14</sup>.

To address the aforementioned limitations of conventional clustering analyses, this study proposed a novel approach for determining markers based on the clustering results and estimation of the following: the rate constants of MPs, biological

<sup>1</sup>Center for Water Cycle Research, Korea Institute of Science and Technology (KIST), Seoul 02792, Republic of Korea. <sup>2</sup>Division of Energy and Environment Technology, KIST-School, University of Science and Technology, Seoul 02792, Republic of Korea. <sup>3</sup>Mass Spectrometer Laboratory, National Instrumentation Center for Environmental Management, Seoul 08826, Republic of Korea. <sup>4</sup>School of Earth Sciences and Environmental Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea. <sup>5</sup>These authors contributed equally: Seung Ji Lim, Jangwon Seo. <sup>6</sup>These authors jointly supervised this work: Moon Son, Seok Won Hong.

✉email: moonson@kist.re.kr; swhong@kist.re.kr



**Fig. 1 Removal efficiency of MPs under aerobic and anoxic conditions.** The sorption and biodegradation in the **a** aerobic and **b** anoxic processes were separately represented. Negative removal is expressed as 0%.

degradation rate ( $k_{bio}$ ), and sorption coefficients ( $K_d$ ); these were determined based on classification and the identified markers. When clusters are established, a marker MP in each cluster can potentially provide information on the behavior, i.e., rate constants, of other MPs in the same cluster. For this purpose, sequential analyses of the self-organizing map (SOM) and random forest classifier (RFC) methods were implemented for analyzing the dataset of 42 MPs consisting of physicochemical properties, functional groups, the initial biotransformation rules, and the rate constants that were obtained under aerobic and anoxic conditions. The SOM, an unsupervised neural network algorithm, was chosen due to its ability for clustering and dimensionality reduction with superior visualization of the input features<sup>15</sup>. In addition, the RFC, a supervised algorithm, was adopted because of its remarkable inference performance for tabular datasets<sup>16,17</sup>. The RFC was utilized to classify the dataset depending on the input features, i.e., the physicochemical properties, functional groups, and the initial biotransformation rules. Hence, the objectives of this study were to: (1) propose the appropriate clustering method for MPs using clustering analysis, (2) determine marker constituents for each cluster aggregated by the SOM, (3) classify MPs using the RFC based on physicochemical properties, functional groups, and biotransformation rules, and (4) estimate a range of rate constants for unlabeled MPs. The results suggest that this approach provides a good framework for monitoring the fate of MPs and can be used as an efficient and effective tool to further reduce the monitoring overheads in WWTPs.

## RESULTS AND DISCUSSION

### Removal of micropollutants under aerobic and anoxic conditions

Figure 1 shows the removal efficiency of 42 MPs after 24 h of incubation. Regardless of the aerobic and anoxic conditions, the removal efficiency of most MPs was less than 5% in the control

experiment. These findings suggest that abiotic processes such as hydrolysis are unlikely to be involved in the transformation of MPs<sup>18,19</sup>. Likewise, volatilization is not considered as the main removal route, because the Henry's constant of the MPs was significantly low<sup>20</sup>. Conversely, adsorption was primarily relevant within 1 h for selected MPs, which is in line with the results of previous studies<sup>21,22</sup>. More than 30% (32–57%) of parabens (methyl paraben, ethyl paraben, propyl paraben, and butyl paraben), estrogens (estrone and estriol), diclofenac, and atorvastatin was removed by sorption onto sludge. However, the removal efficiency of most MPs through sorption was less than 14% (Fig. 1).

Distinct differences were observed in the biodegradation of each MPs. For example, ibuprofen, naproxen, caffeine, metformin, gemfibrozil, and acetaminophen were almost completely removed under aerobic conditions (Fig. 1a). The removal of these MPs primarily resulted from biodegradation, which is consistent with previous findings<sup>21,23,24</sup>. Atorvastatin, parabens, and estrogens were also completely removed (Fig. 1a) through sorption and biodegradation, accounting for 36–70% and 30–64% of the removal, respectively. On the other hand, antibiotics (sulfathiazole, sulfamethazine, sulfamethoxazole, trimethoprim, and lincomycin), carbamazepine, atrazine, clofibric acid, and *N,N*-diethyl-metolamide were poorly removed (sorption: less than 5% and biodegradation: up to 40%). Less removal of these MPs is consistent with the findings of Ternes, et al.<sup>20</sup> and Joss, et al.<sup>25</sup>; this is mainly ascribed to less reactivity of the functional groups<sup>26</sup>.

Propranolol exhibited negative removal under aerobic conditions (Fig. 1), most likely due to back-transformation of propranolol and deconjugation<sup>27</sup>. Within 24 h, approximately 70–85% of atenolol, ranitidine, iopromide, cimetidine, and gemfibrozil were removed mainly by biodegradation. The total removal efficiency of diclofenac was approximately 74% (Fig. 1a). However, sorption (47%) was more effective than biodegradation (27%) for the removal of diclofenac. This is in line with the previous studies demonstrating the sorption onto sludge is a

predominant route of diclofenac removal<sup>22,28,29</sup>. Perfluorinated compounds were removed less efficiently (7.5–37.3%), except for perfluoropentanoic acid (negatively removed). The biological removal efficiencies of *N*-nitrosamines varied significantly, as shown in Fig. 1a, with *N*-nitrosomorpholine (5.7%) and *N*-nitrosodibutylamine (75%) exhibiting the lowest and highest removal rates, respectively. The wide range of total removal efficiencies of 12.6–79.4% (sorption: less than 6% and biodegradation: 5.7–75%) in this study is analogous to the results of previous studies (>10–94%)<sup>30–32</sup>. Nitrosamines with acyclic groups, such as *N*-nitrosodibutylamine, were more biodegradable than those with alicyclic and morpholine groups<sup>31</sup>.

Metformin was almost completely biodegraded under anoxic conditions, whereas MPs such as ranitidine, iopromide, and acetaminophen were removed with varying degrees of efficiency (62–85%). Also, parabens, estriol, and estrone were significantly removed under anoxic conditions, most likely due to sorption (Fig. 1b). Biodegradation efficiency of  $\beta$ -blockers (atenolol, metoprolol, and propranolol) and trimethoprim was 45% higher under anoxic conditions than that under aerobic conditions. In contrast, corrosion inhibitors (1-*H*-benzotriazole and 4-methyl-1-*H*-benzotriazole), gemfibrozil, diclofenac, ibuprofen, naproxen, and caffeine were only marginally removed (7.9–28.6%). Alvarino et al.<sup>33</sup> and Mazioti et al.<sup>34</sup> reported that the degradation of these MPs was only observed under aerobic conditions depending on the activity of nitrifying bacteria. Perfluoropentanoic acid had a negative removal efficiency in this study, while other MPs had less than 35% removal efficiency via biodegradation under anoxic conditions (Fig. 1b). The negative removal of perfluoropentanoic acid could be resulted from the transformation of other perfluorinated compounds present in the sludge samples<sup>35</sup>.

Overall, most results of batch experiments under both conditions were consistent with previously reported results. According to the Nash–Sutcliffe model efficiency coefficient (NSE) results, the pseudo first-order degradation model can predict the majority of biodegradation rate constants of MP under aerobic and anoxic conditions except for perfluoropentanoic acid, perfluorobutanesulfonate, atrazine, and nitrosamines. The  $k_{bio}$  of MPs ranged from 0 to 2.3 L g<sup>-1</sup> h<sup>-1</sup> and 0 to 1.8 L g<sup>-1</sup> h<sup>-1</sup> under aerobic and anoxic conditions, respectively (Table 1). As shown in Supplementary Fig. 1, however, most of the MPs showed higher  $k_{bio}$  values in the aerobic process than those in the anoxic process. The sorption coefficients were very similar under both aerobic (0–0.44 L g<sub>MLSS</sub><sup>-1</sup>) and anoxic conditions (0–0.5 L g<sub>MLSS</sub><sup>-1</sup>), which agreed well with previously reported ranges<sup>25,28,36–40</sup>.

### Machine learning models combining clustering and classification

Machine learning models that combine clustering and classification are proposed in this study (Fig. 2). The clustering analysis was used to assign a label to unlabeled data that could be further used during classification<sup>41</sup>. The dataset consisted of physicochemical properties, functional groups, initial biotransformation rules, and rate constants of 42 MPs (Supplementary Table 1). The dataset was randomly divided into two parts: 29 MPs for the training and validation (70% for cross-validation) datasets and 13 MPs for the test (30%) dataset. It is noted that the abbreviations are used to indicate clustering scenarios based on the physicochemical properties and functional groups as PF and the initial biotransformation rules as BT.

### Clustering analysis and determination of marker constituents

The distance maps derived from the SOM are illustrated with different colors according to the relative distance between each neuron of the map (Figs. 3a and 4a). The MPs assigned closely in the distance map filled with similar colors were interpreted as MPs having analogous characteristics (Figs. 3b and 4b). The radius of

sectors represents the relative importance of each input feature to cluster MPs. On the other hand, the MPs with remarkably different features were in the separate neurons with dissimilar colors. The solid lines determined by Ward's method indicate the boundaries dividing each cluster. The marker constituents among MPs are indicated with superscripts (A) for aerobic and (AN) for anoxic conditions, respectively (Figs. 3 and 4).

### Clustering analysis based on physicochemical properties and functional groups

Recent research has found that the fate of MPs is influenced by physicochemical properties such as the octanol-water partition coefficient and accessible functional moieties<sup>10,20,26</sup>. Hence, we first assessed the suitability of physicochemical properties and functional groups for clustering MPs (Fig. 3). Using Ward's method, MPs having similar input features were clustered into 11 clusters with the lowest Davies–Bouldin index (DBI) in the PF scenario (0.49).

Because nitrosamines commonly contain amine and amide functional groups, they are clustered together as shown in the left upper side of Fig. 3a, b. However, *N*-nitrodiphenylamine and *N*-nitrosomorpholine were assigned in different clusters due to having diphenylamine and morpholine as aromatic functional group, respectively. Carbamazepine and *N,N*-diethyl-meta-toluamide were also grouped together with *N*-nitrodiphenylamine and *N*-nitrosomorpholine because they contain amine, amide, and aromatic ring as functional groups. The MPs having nitrogen- and sulfur-containing functional groups such as sulfathiazole, sulfamethazine, ranitidine, and cimetidine were assigned to one cluster in the lower-left corner of Fig. 3a. This clustering result is line with previous studies in which MPs with sulfonamide functional group were aggregated in the same cluster<sup>14</sup> and sulfamethazine and sulfathiazole were closely located in the dendrogram on the basis of biodegradation rate<sup>11</sup>. The parabens were clustered in the same unit because of their high log  $K_{ow}$  values and functional groups, i.e., ester and aromatic ring. The long alkyl ester chain and high log  $K_{ow}$  value are the unique properties of parabens, which lead readily to sorption and biodegradation<sup>42</sup>. Although estrogens do not have an ester functional group in their structure, parabens and estrogens were assigned in the same cluster due to their similarity in log  $K_{ow}$  value and having alcohol and aromatic ring as functional groups (upper right corner of Fig. 3a, b). The MPs located in the lower-right corner of Fig. 3a, b contain a halogen-containing functional group in common. The perfluorohexanoic acid and perfluoropentanoic acid were separately clustered from clofibric acid and iopromide because of the fluorinated carbon chain in their structure rather than the aromatic ring structure. Similar clustering results can be found in the previous study in which perfluorinated compounds were grouped in the same cluster due to their fluorinated carbon chain structure<sup>43</sup>.

In summary, the clustering result represented in the SOM map (Fig. 3a) was interpretable using the physicochemical properties of each MP (Fig. 3b). In the figure, the MPs on the left side have relatively low molecular weights or log  $K_{ow}$  and consist of nitrogen-containing functional groups (i.e., amine and amide) compared with the MPs on the right side. The MPs having the aromatic ring functional group were in a diagonal direction (lower left to upper right), and the MPs with the chain structure were positioned at each corner, in the upper left and lower right. Other MPs containing sulfur and halogen atoms in their functional groups aggregated in the clusters at the bottom of the distance map. One limitation of the clustering result in this study was the uneven distribution of MPs in each cluster due to the lack of available MP data. This limitation should be overcome in future studies by increasing the number of MPs included in the analysis.

**Table 1.** Sorption coefficient and biological reaction rate constant of MPs in aerobic and anoxic processes estimated through the dynamic modeling approach.

Category	Micropollutant	$K_d, aerobic$ (L g <sub>M</sub> LSS <sup>-1</sup> )	$k_{bio, aerobic}$ (L g <sup>-1</sup> h <sup>-1</sup> )	NSE <sub>aerobic</sub>	$K_d, anoxic$ (L g <sub>M</sub> LSS <sup>-1</sup> )	$k_{bio, anoxic}$ (L g <sup>-1</sup> h <sup>-1</sup> )	NSE <sub>anoxic</sub>
Antidiabetic	Metformin	0.008 ± 0.012	0.122 ± 0.047	0.949	0.064 ± 0.009	0.086 ± 0.003	0.952
β-blocker	Atenolol	0.019 ± 0.002	0.038 ± 0.007	0.895	0.033 ± 0.047	0.077 ± 0.002	0.940
	Metoprolol	0.007 ± 0.011	0.004 ± 0.001	0.118	0.015 ± 0.004	0.007 ± 0.000	0.647
	Propranolol	0.05 ± 0.024	0.000 ± 0.002	-0.041	0.024 ± 0.014	0.013 ± 0.001	0.276
H <sub>2</sub> -receptor blocker	Cimetidine	0.018 ± 0.007	0.024 ± 0.003	0.901	0.012 ± 0.003	0.005 ± 0.000	0.501
	Ranitidine	0.178 ± 0.001	0.087 ± 0.003	0.773	0.126 ± 0.029	0.019 ± 0.001	0.932
Antibiotic	Lincomycin	0.015 ± 0.001	0.004 ± 0.000	0.553	0.019 ± 0.016	0.002 ± 0.000	0.614
	Sulfathiazole	0.038 ± 0.006	0.008 ± 0.003	0.929	0.060 ± 0.031	0.003 ± 0.001	0.817
	Sulfamethazine	0.012 ± 0.009	0.006 ± 0.002	0.834	0.023 ± 0.002	0.002 ± 0.001	0.677
	Sulfamethoxazole	0.034 ± 0.005	0.007 ± 0.002	0.936	0.044 ± 0.007	0.001 ± 0.000	0.403
	Trimethoprim	0.064 ± 0.006	0.004 ± 0.001	-0.662	0.088 ± 0.023	0.045 ± 0.001	0.944
Non-steroidal anti-inflammatory drug	Acetaminophen	0.036 ± 0.012	0.165 ± 0.001	0.945	0.018 ± 0.001	0.048 ± 0.001	0.913
	Diclofenac	0.303 ± 0.026	0.039 ± 0.008	0.603	0.300 ± 0.139	0.009 ± 0.002	0.619
	Ibuprofen	0.052 ± 0.002	0.276 ± 0.033	0.853	0.022 ± 0.014	0.008 ± 0.001	0.591
	Naproxen	0.023 ± 0.007	0.103 ± 0.034	0.760	0.050 ± 0.004	0.006 ± 0.001	0.655
Anticonvulsant	Carbamazepine	0.046 ± 0.004	0.000 ± 0.001	-0.264	0.032 ± 0.001	0.002 ± 0.001	0.717
Anti-dyslipidemia agent	Atorvastatin	0.187 ± 0.071	0.142 ± 0.020	0.225	0.128 ± 0.417	0.018 ± 0.004	0.575
	Clofibrilic acid	0.010 ± 0.012	0.002 ± 0.001	0.670	0.000 ± 0.003	0.003 ± 0.001	-1.657
	Gemfibrozil	0.065 ± 0.007	0.056 ± 0.016	0.883	0.040 ± 0.025	0.003 ± 0.002	0.568
X-ray contrast agent	Iopromide	0.051 ± 0.017	0.048 ± 0.005	0.953	0.027 ± 0.012	0.025 ± 0.002	0.431
Corrosion inhibitor	1-H-benzotriazole	0.010 ± 0.021	0.019 ± 0.001	0.839	0.004 ± 0.004	0.005 ± 0.001	0.503
	4-methyl-1H-benzotriazole	0.034 ± 0.002	0.005 ± 0.000	0.759	0.041 ± 0.001	0.002 ± 0.001	0.627
Others	Caffeine	0.046 ± 0.015	0.165 ± 0.060	0.992	0.032 ± 0.005	0.003 ± 0.001	0.727
Preservative	Methyl paraben	0.279 ± 0.001	0.132 ± 0.008	0.847	0.187 ± 0.369	0.140 ± 0.037	0.923
	Ethyl paraben	0.155 ± 0.040	1.868 ± 0.064	0.762	0.208 ± 0.267	1.549 ± 0.043	0.957
	Propyl paraben	0.210 ± 0.044	1.592 ± 0.367	0.831	0.197 ± 0.228	1.195 ± 0.244	0.916
	Butyl paraben	0.439 ± 0.131	2.295 ± 0.109	0.811	0.266 ± 0.100	1.850 ± 0.051	0.897
Endocrine disruptor	Estrilol	0.349 ± 0.102	1.852 ± 0.409	0.888	0.505 ± 0.054	1.094 ± 0.422	0.925
	Estrone	0.226 ± 0.327	0.169 ± 0.041	0.798	0.070 ± 0.402	0.029 ± 0.008	0.614
Pesticide	Atrazine	0.004 ± 0.014	0.001 ± 0.000	-0.569	0.007 ± 0.027	0.001 ± 0.000	-0.951
Biocide	N,N-diethyl-meta-toluamide	0.009 ± 0.022	0.003 ± 0.001	0.789	0.026 ± 0.009	0.001 ± 0.000	0.164
Perfluorinated compound	Perfluoropentanoic acid	0.004 ± 0.007	0.000 ± 0.001	-1.259	0.020 ± 0.012	0.000 ± 0.001	-0.192
	Perfluorobutanesulfonic acid	0.009 ± 0.015	0.004 ± 0.001	-2.229	0.035 ± 0.121	0.002 ± 0.001	-0.172
	Perfluorohexanoic acid	0.001 ± 0.003	0.001 ± 0.001	0.615	0.005 ± 0.004	0.000 ± 0.001	-0.583
	Perfluoroheptanoic acid	0.000 ± 0.011	0.003 ± 0.002	0.575	0.039 ± 0.022	0.001 ± 0.001	-0.403
Nitrosamine	N-nitrosodimethylamine	0.023 ± 0.080	0.001 ± 0.001	-2.020	0.070 ± 0.096	0.001 ± 0.001	-0.499
	N-nitrosoethylmethylamine	0.023 ± 0.077	0.001 ± 0.001	-2.562	0.071 ± 0.100	0.001 ± 0.001	-0.251
	N-nitrosodiethylamine	0.023 ± 0.078	0.002 ± 0.001	-2.330	0.053 ± 0.086	0.001 ± 0.001	-0.275
	N-nitrodiphenylamine	0.018 ± 0.062	0.006 ± 0.002	-2.436	0.004 ± 0.029	0.000 ± 0.001	-0.185
	N-nitrosopyrrolidine	0.019 ± 0.095	0.002 ± 0.001	-2.195	0.082 ± 0.111	0.006 ± 0.001	-0.088
	N-nitrosomorpholine	0.025 ± 0.079	0.001 ± 0.001	-1.246	0.033 ± 0.069	0.001 ± 0.001	-0.219
	N-nitrosodibutylamine	0.016 ± 0.103	0.033 ± 0.002	0.743	0.014 ± 0.002	0.002 ± 0.002	-0.066

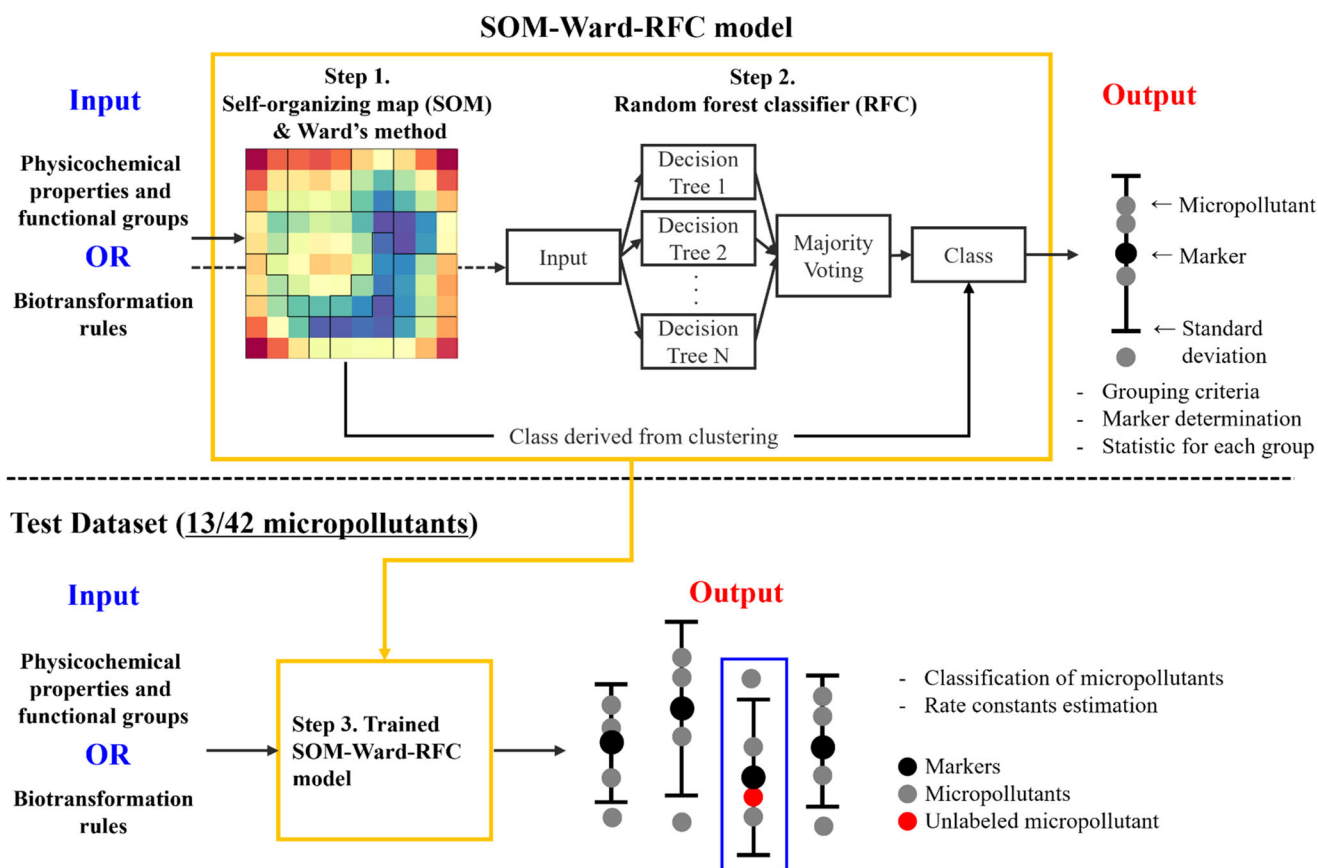
### Clustering analysis based on biotransformation rules

Perfluorinated compounds, N-nitrosodimethylamine, and N-nitrosopyrrolidine were excluded in this section because their initial biotransformation rules were not predictable using Eawag-PPS. When the SOM clustered the MPs based on the BT scenario, the algorithm generated 15 clusters (DBI = 0.87). The MPs, most commonly following the 1–3 initial biotransformation rules, were aggregated in the same cluster (Fig. 4). 1-H-benzotriazole was grouped together with clofibrilic acid because its biotransformation

was mainly initiated by the aromatic ring dihydroxylation (bt0005) (right middle area of Fig. 4a, b). Atenolol and iopromide were aggregated in the same cluster, since the biodegradation of atenolol and iopromide likely occurred through H-abstraction from side chains (bt0002) and demethylation or dealkylation of ether group (bt0023) (lower-left area of Fig. 4a, b). These results were not consistent with the previous report demonstrating that atenolol and 1-H-benzotriazole were tied to the same cluster when using the elimination rates instead of the biotransformation rule as input features<sup>43</sup>. Since sulfathiazole and sulfamethazine



## Training and Validation Datasets (29/42 micropollutants)



**Fig. 2 Schematic diagram of the combination of clustering and classification algorithms.** The number of micropollutants (MPs) used in each step is noted in the diagram.

contain a sulfonamide functional group, which can be biodegraded through hydrolysis and bond-cleavage in sulfonamide group (bt0144), they were aggregated together in the same cluster. This is in line with a previous study showing that the MPs having the sulfonamide functional group were aggregated in the same cluster<sup>10,14</sup>. Among the nitrosamine compounds, *N*-nitrosodiethylamine and *N*-nitrosomethylethylamine, biodegradation mainly resulted from the monohydroxylation of methyl group (bt0334), and hence were aggregated in one cluster.

### Estimation of rate constants using the proposed algorithms and markers

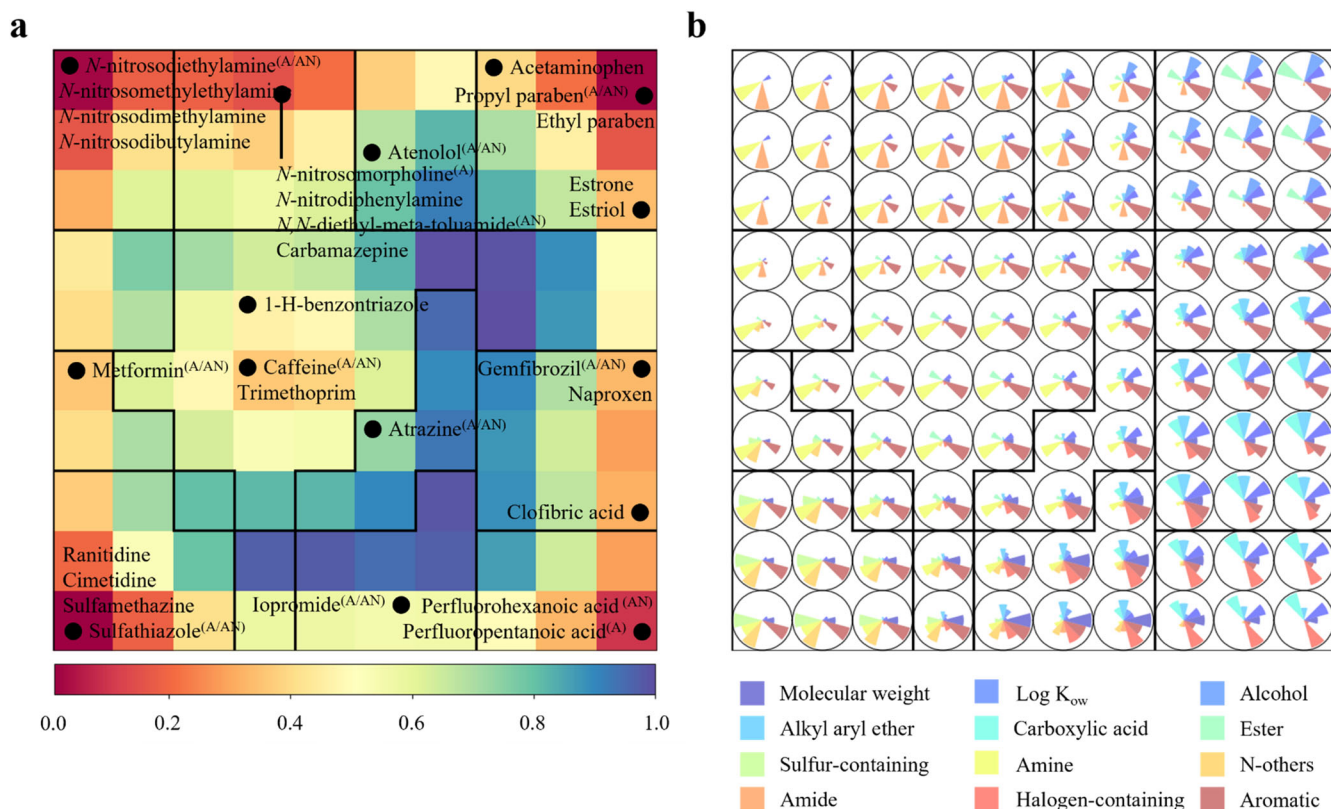
The feasibility of the proposed algorithms and derived marker constituents was evaluated by classifying unlabeled MPs, followed by estimating the range of rate constants for each MP. In this study, the role of a marker is to provide representative information regarding the rate constants of MPs in each cluster. Therefore, the marker was designated as an MP having a minimum Euclidean distance from the mean of the rate constants in each cluster. For example, when an unlabeled MP is classified in a specific cluster, the ranges of its rate constants can be calculated using the rate constants of the markers,  $K_{d,m}$  and  $k_{bio,m}$ . The rate constants for unlabeled MPs,  $K_{d,u}$  and  $k_{bio,u}$ , can have the values in the range as follows:

$$K_{d,m} - N\sigma_{K_d} \leq K_{d,u} \leq K_{d,m} + N\sigma_{K_d} \quad (1)$$

$$k_{bio,m} - N\sigma_{k_{bio}} \leq k_{bio,u} \leq k_{bio,m} + N\sigma_{k_{bio}} \quad (2)$$

where  $\sigma_{K_d}$  and  $\sigma_{k_{bio}}$  indicate the standard deviation of sorption coefficient and biodegradation rate constant obtained from the MPs in each cluster, respectively.  $N$  was set to one, two, and three in this study. The estimation accuracy was calculated by counting the numbers of MPs that lie within the range calculated using Eqs. (1) and (2) (Table 2).

In the preliminary simulation to design this study, a random forest regressor (RFR) was solely employed to directly predict the degradation rate constants (Supplementary Fig. 2). The coefficient of determination ( $R^2$ ) for degradation rate constants in the test step was lower than 0.5 regardless of input features and operating conditions (Supplementary Figs. 3 and 4). An overfitting problem that the prediction accuracy for the training step was significantly higher ( $R^2$ : 0.78–0.90) than the test step ( $R^2$ : –0.08–0.45) occurred in the RFR model. However, the machine learning approach combining SOM and RFC performed better than the RFR model only; hence, the SOM and RFC were utilized in this study. In the training and validation steps, the classification accuracy (0.75) and f1-score (0.61) of the PF scenario were significantly higher than those of the BT scenario (accuracy: 0.43 and f1-score: 0.32). In the test step, with respect to the aerobic condition, the algorithm using the PF scenario was able to estimate the range for rate constants with an accuracy of 0.38 using one standard deviation and marker's rate constants of each cluster. In contrast, one standard deviation was insufficient to estimate the range of rate constants in the BT scenario (0.10). The best estimation accuracy of the BT scenario (0.40) was relatively lower than that of the PF scenario when the estimations were made within three standard deviations (0.77). Similar to the aerobic condition, under the



**Fig. 3 Clustering results for physicochemical properties and functional groups of MPs. a** Distance between neighboring map units and clustering boundary and **b** weight vectors that represent the importance of each feature to organize the map. The color bar indicates the distance between neighboring map units.

anoxic condition, the estimation accuracy for the PF scenario (0.46–0.77) showed better estimation performance compared with that of the BT scenario (0.30–0.40). Collectively, the PF scenario showed higher performance in the classification of MPs and estimation of rate constants compared with the BT scenario. The better clustering results can explain this relatively higher classification and estimation accuracies of the PF scenario than the BT scenario. For example, the DBI value for the PF scenario (0.49) was only half of the DBI value for the BT scenario (0.87), implying that the clustering using the PF scenario was more well-organized than that of the BT scenario.

#### Applicability of the proposed algorithm to different microbial community data

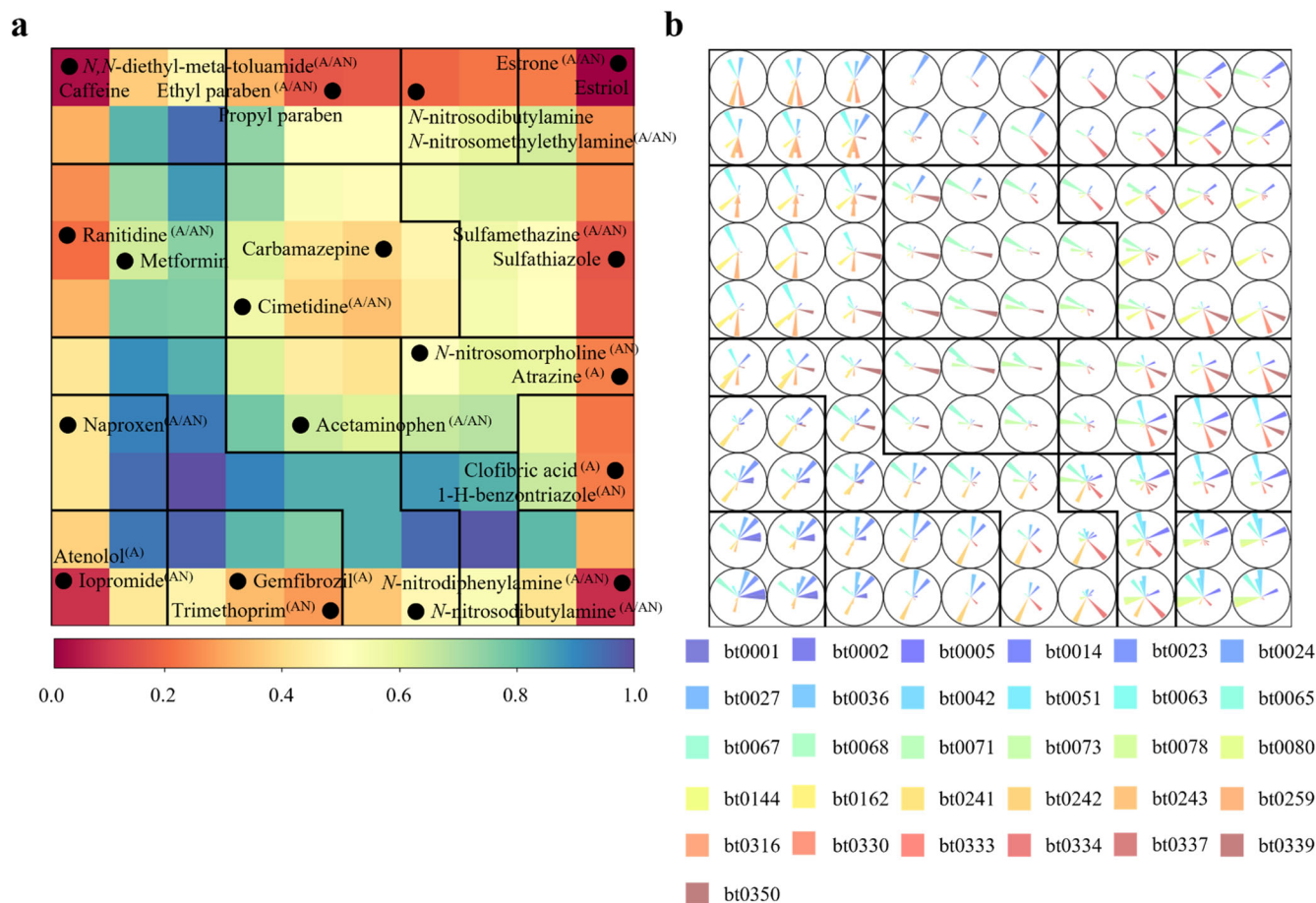
We further conducted simulations using the previously reported aerobic experimental data to evaluate the applicability of this proposed machine learning algorithm to different microbial community data<sup>10</sup>. In this case, the dataset consisted of 42 MPs, mainly degraded through biotransformations but not sorptions. The proposed algorithm was retrained using physicochemical properties, functional groups, and biotransformation to estimate the rate constants of MPs in the reported datasets. As expected, the proposed algorithm was able to classify MPs and estimate the rate constants of MPs in the different microbial community. Interestingly, in this case, the BT scenario (0.72) showed a slightly higher classification accuracy than that of the PF scenario (0.62). Consequently, the estimation performance using the BT scenario (0.69) was also slightly higher than that of the PF scenario (0.62) (Supplementary Table 2). This superior estimation accuracy under the BT scenario is likely because the selected MPs in the literature datasets follow the rules of biotransformation well, as stated<sup>10</sup>. However, the use of the biotransformation rules only as input

features led to a lower estimation performance of rate constants compared with the PF scenario for our experimental datasets. This can be ascribed to the fact that the sorption is indirectly counted under the PF scenario, which has considered the physicochemical properties and functional groups of MPs but not under the BT scenario. As a result, the estimation of rate constants could not be precisely conducted under the BT scenario.

#### Comparison of model performances with previous studies

The proposed algorithm exhibited a comparable classification performance and superior estimation accuracy of MPs when estimating the range of rate constants compared to the ones proposed by previous studies. For example, a previous model based on meta-analysis accounted for only 17% of the variability in the removal efficiencies of the targeted MPs<sup>44</sup>, which is lower than the performance of the PF scenario under the aerobic condition within one standard deviation (0.38). In another study employing hierarchical clustering and multivariable analysis, the estimation accuracy for the complete dataset was only 0.19 owing to the unpredictable characteristics of biodegradation<sup>14</sup>. A recent study proposed an RFC to classify MPs into two classes (fast or slow biotransformation) with classification accuracies of 0.95 for the predicted biotransformation rules and 0.78 for the observed biotransformation rules<sup>45</sup>. This classification accuracy is similar to the present study. Importantly, in this study, a direct estimation of the range of rate constants of unlabeled MPs was possible. However, the previous study could only classify whether the MPs were biodegraded slowly or rapidly.

Overall, the superior estimation accuracy of this proposed machine learning algorithm suggested two noteworthy findings. First, the markers represented each cluster successfully, particularly when the physicochemical properties and functional groups



**Fig. 4** Clustering results for initial biotransformation rules of MPs. **a** Distance between neighboring map units and clustering boundary and **b** weight vectors. The initial biotransformation rules for each MP are presented in Supplementary Table 1. The color bar indicates the distance between neighboring map units.

of each MP were employed during the model training. Second, the markers derived from the proposed algorithm were used to estimate the range of rate constants for unlabeled MPs in the test dataset with relatively high accuracy, using only their physicochemical properties and functional groups as input features. In summary, the proposed machine learning approach could be employed to estimate the sorption and degradation rate of unlabeled and emerging MPs based only on the physicochemical properties and functional groups rather than measuring time-course change of their concentration to estimate the fate of MPs. The proposed machine learning approach trained with sufficient process operational and experimental data could reduce the labor and expenses required for monitoring MPs. Thus, monitoring only the marker MP could reduce the cost of measuring each MP concentration. As with other machine learning techniques, one important prerequisite for successfully applying this machine learning model is to secure sufficient data to train the model. With sufficient data, the grouping and positioning of MPs with SOM could become more refined while improving the accuracy of predictions with RFCs.

## METHODS

The details of the activated sludge, reagents, and chemicals used in this study are provided in the Supplementary Information (See Supplementary Note 1). Unless otherwise noted, all experiments were conducted using synthetic wastewater (SyWW). The detailed composition of SyWW is presented in Supplementary Table 3.

## Batch experiments

The biodegradation of 42 MPs was evaluated under aerobic and anoxic redox conditions. These MPs were chosen because of their frequency of occurrence, persistence, and negative impact on aquatic life. The agitated batch reactor setups are presented in Supplementary Note 2 and Supplementary Fig. 5. Approximately 2.2 L SyWW with 0.8 L activated sludge was filled in 3 L batch reactors. A cocktail of 42 MPs was spiked into the reactors with a final concentration of  $0.1 \text{ mg L}^{-1}$ . The concentration values of mixed liquor suspended solids (MLSS) and mixed liquor volatile suspended solids (MLVSS) were maintained at  $3 \text{ g L}^{-1}$  and  $1.8 \text{ g L}^{-1}$  in all the experiments, respectively. The pH and the water temperature were kept at  $\text{pH} = 7$  and  $22 \text{ }^\circ\text{C}$ , respectively, throughout the experiment. An 11 mL aliquot of the sample solution was collected from the reactors at the following periods: 0, 10, 20, and 30 min and 1, 2, 4, 8, 12, and 24 h. Control experiments without sludge were also performed to verify abiotic transformation of MPs with a sampling interval of 0 and 24 h. To investigate the adsorption effect on their removal (sterile control), the samples from the reactor spiked with sodium azide ( $3 \text{ g L}^{-1}$ ) to suppress the microbial activity were collected at 0, 10, 20, 30 min, and 1 h. Prior to analysis, all samples were filtered using a  $0.2 \text{ } \mu\text{m}$  syringe filter (Whatman), fortified with internal standards ( $50 \text{ ng mL}^{-1}$ ), and immediately stored in a freezer at  $-20 \text{ }^\circ\text{C}$ .

## Analysis of micropollutants

Nitrosamines were analyzed using gas chromatography coupled to low-resolution mass spectrometry (GC-LRMS, 6890 N GC system,



**Table 2.** Performance of clustering (DBI) and classification (accuracy, f1-score, precision, and recall) approaches and estimation performance of rate constants range using markers for each clustering scenario.

Process	Input feature	DBI	Accuracy	f1-score	Precision	Recall	Estimation accuracy ( $N = 1$ )	Estimation accuracy ( $N = 2$ )	Estimation accuracy ( $N = 3$ )
Aerobic	Physicochemical properties and functional groups	0.488	0.75 (0.15)	0.61 (0.24)	0.59 (0.24)	0.65 (0.23)	0.38	0.69	0.77
	Biotransformation rule	0.872	0.43 (0.16)	0.32 (0.13)	0.31 (0.14)	0.33 (0.11)	0.10	0.20	0.40
Anoxic	Physicochemical properties and functional groups	0.488	0.75 (0.15)	0.61 (0.24)	0.59 (0.24)	0.65 (0.23)	0.46	0.70	0.77
	Biotransformation rule	0.872	0.43 (0.16)	0.32 (0.13)	0.31 (0.14)	0.33 (0.11)	0.30	0.40	0.40

The standard deviation of classification performance is mentioned in the parentheses. The estimation performance was compared with different  $N$  values, deciding the estimation range of rate constants in Eqs. (1) and (2).

Agilent Technologies, USA). The details of the procedure and validation of the GC-LRMS method have been reported elsewhere<sup>46</sup>. Thirty-five additional MPs were monitored using an ultra-high performance liquid chromatography (UHPLC) Vanquish system (Thermo Scientific, San Jose, USA). The system consisted of a cooling auto-sampler, column oven enabling temperature control, ultra-high pressure solvent delivery pump, and automatic degasser. Chromatographic separations of the samples were performed using a Cortecs C18 column (100 × 2.1 mm, 1.6 μm, Waters Co., Milford, MA, USA). The column temperature was set at 45 °C, and the injection volume was 3 μL with a flow rate of 0.3 mL min<sup>-1</sup>. The mobile phases included 0.1% hydrofluoric acid in high-performance liquid chromatography (HPLC) grade water (Solvent A) and methanol (Solvent B). The gradient elution consisted of 0–0.5 min, 40–70% B, 0.5–6.5 min, 70–100% B and a 1 min hold time, followed by a 4 min re-equilibration to the starting conditions. The internal standards were used for quantification of analytes and a ten-point calibration curve was constructed with a concentration range of 0.1 to 100 ng mL<sup>-1</sup>. Details of the optimization and validation of the UHPLC-MS/MS methods are described in Supplementary Note 3, Supplementary Tables 4 and 5.

### Pseudo first-order degradation models

Based on the results obtained from lab-scale batch experiments, a pseudo first-order degradation kinetic model (Supplementary Note 4) has been frequently used for describing the fate of MPs<sup>25,38,47</sup>. The pseudo first-order degradation model in this study assumed fast sorption that reached the equilibrium condition immediately due to observation of instant reduction of soluble MP concentration. Other degradation/removal mechanisms such as volatilization were not considered. The performance of the model was evaluated using the NSE (Supplementary Note 5). Within the scope of this study, the pseudo first-order degradation model considering  $k_{bio}$  and  $K_d$  can effectively describe the kinetics of MPs.

### Machine learning approaches using clustering and classification for micropollutants

In Step 1, the SOM, followed by Ward's method, was employed in the training and validation datasets to cluster MPs in the reduced dimension, mapping high-dimensional data onto a two-dimensional grid. Ward's method draws the decision boundary to effectively separate clusters generated by SOM (Supplementary Note 6). The optimum number of clusters was calculated by evaluating the Davies–Bouldin index (DBI) (Supplementary Note 7). Step 1 aims to assign a label to MPs whose appropriate grouping rules do not yet exist. The MPs in the same cluster were considered to have similar functional groups or biodegradation

rules. The labels derived in this step were used to train the classification algorithm in Step 2. The marker for each cluster was determined after verifying the number of clusters having a minimum DBI. The marker MPs are the representative MP of each cluster, which were used in Step 3 to estimate the degradation rate constants of the unlabeled MP in the test dataset. Two clustering scenarios were designed to find the proper input features for clustering MPs: clustering based on the physicochemical properties and functional groups (e.g., octanol-water partitioning coefficient, ether, ester, and amine functional groups) (see more information provided in Supplementary Table 1) and the initial biotransformation rules predicted from Eawag-PPS as presented in Supplementary Table 1<sup>12,13</sup>.

In Step 2, the RFC was used to establish a classification algorithm predicting labels assigned to the training and validation datasets in the clustering (in Step 1). The input features, i.e., the physicochemical properties, functional groups, and the initial biotransformation rules, used in clustering were also employed to classify MPs to each label. Cross-validation with a five-fold size was conducted to evaluate the classification performance. The clustering scenario with the better classification accuracy and f1-score (Supplementary Note 9) was chosen as the best clustering scenario for the machine learning model.

In Step 3, the trained model (trained SOM-WARD-RFC model in Fig. 2) was utilized to classify the unlabeled MPs. When the trained model classified the unlabeled MP in the test dataset to the established cluster in Step 1, the classified MP could be considered to have similar degradation properties to other MPs in the same cluster. The markers in each cluster were used to estimate the range of rate constants for unlabeled MPs using Eqs. (1) and (2). Since the unlabeled MPs in the test dataset were completely separated from the MPs in the train and validation dataset, there was no possibility that the model had previewed the data used in the test step. The specific operation conditions regarding the SOM and RFC mentioned in this section are given in Supplementary Note 6–9. In this study, all simulations were performed using Python 3.7 and the clustering was conducted using the SOM from *MiniSOM toolbox* version 2.3.0<sup>48</sup>. Ward's method and the RFC from *Scikit-learn* version 1.0 were used to draw decision boundaries and classify MPs depending on input features, respectively<sup>49</sup>.

### DATA AVAILABILITY

All data are available in the manuscript or the supplementary information.

### CODE AVAILABILITY

The underlying code for this study is not publicly available for proprietary reasons.



Received: 2 January 2023; Accepted: 16 October 2023;  
Published online: 28 October 2023

## REFERENCES

- Wang, Z., Walker, G. W., Muir, D. C. G. & Nagatani-Yoshida, K. Toward a global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories. *Environ. Sci. Technol.* **54**, 2575–2584 (2020).
- Patel, M. et al. Pharmaceuticals of emerging concern in aquatic systems: chemistry, occurrence, effects, and removal methods. *Chem. Rev.* **119**, 3510–3673 (2019).
- EGgen, R. I. L., Hollender, J., Joss, A., Schäfer, M. & Stamm, C. Reducing the discharge of micropollutants in the aquatic environment: the benefits of upgrading wastewater treatment plants. *Environ. Sci. Technol.* **48**, 7683–7689 (2014).
- Luo, Y. et al. A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. *Sci. Total Environ.* **473–474**, 619–641 (2014).
- Rout, P. R., Zhang, T. C., Bhunia, P. & Surampalli, R. Y. Treatment technologies for emerging contaminants in wastewater treatment plants: a review. *Sci. Total Environ.* **753**, 141990 (2021).
- Buerge, I. J., Kahle, M., Buser, H. R., Müller, M. D. & Poiger, T. Nicotine derivatives in wastewater and surface waters: application as chemical markers for domestic wastewater. *Environ. Sci. Technol.* **42**, 6354–6360 (2008).
- Tran, N. H., Li, J., Hu, J. & Ong, S. L. Occurrence and suitability of pharmaceuticals and personal care products as molecular markers for raw wastewater contamination in surface water and groundwater. *Environ. Sci. Pollut. Res.* **21**, 4727–4740 (2014).
- Buerge, I. J., Poiger, T., Müller, M. D. & Buser, H. R. Caffeine, an anthropogenic marker for wastewater contamination of surface waters. *Environ. Sci. Technol.* **37**, 691–700 (2003).
- Buerge, I. J., Poiger, T., Müller, M. D. & Buser, H. R. Combined sewer overflows to surface waters detected by the anthropogenic marker caffeine. *Environ. Sci. Technol.* **40**, 4096–4102 (2006).
- Achermann, S. et al. Trends in micropollutant biotransformation along a solids Retention time gradient. *Environ. Sci. Technol.* **52**, 11601–11611 (2018).
- Desiante, W. L., Minas, N. S. & Fenner, K. Micropollutant biotransformation and bioaccumulation in natural stream biofilms. *Water Res.* **193**, 116846 (2021).
- Ellis, L. B. & Wackett, L. P. Use of the University of Minnesota Biocatalysis/Biodegradation Database for study of microbial degradation. *Micro. Inf. Exp.* **2**, 1 (2012).
- Ellis, L. B., Gao, J., Fenner, K. & Wackett, L. P. The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res.* **36**, W427–W432 (2008).
- Wang, Y., Fenner, K. & Helbling, D. E. Clustering micropollutants based on initial biotransformations for improved prediction of micropollutant removal during conventional activated sludge treatment. *Environ. Sci. Water Res. Technol.* **6**, 554–565 (2020).
- Kohonen, T. Essentials of the self-organizing map. *Neural Netw.* **37**, 52–65 (2013).
- Shwartz-Ziv, R. & Armon, A. Tabular data: deep learning is not all you need. *Inf. Fusion* **81**, 84–90 (2022).
- Ullah, Z., Yoon, N., Tarus, B. K., Park, S. & Son, M. Comparison of tree-based model with deep learning model in predicting effluent pH and concentration by capacitive deionization. *Desalination* **558**, 116614 (2023).
- Williams, M., Du, J., Kookana, R. & Azzi, M. In *Biodegradation, hydrolysis and photolysis testing of nitrosamines in aquatic systems*, 1–30 (Commonwealth Scientific and Industrial Research Organisation, 2011).
- Bergheim, M., Gieré, R. & Kümmerer, K. Biodegradability and ecotoxicity of tramadol, ranitidine, and their photoderivatives in the aquatic environment. *Environ. Sci. Pollut. Res.* **19**, 72–85 (2012).
- Ternes, T. A. et al. A rapid method to measure the solid-water distribution coefficient ( $K_d$ ) for pharmaceuticals and musk fragrances in sewage sludge. *Water Res.* **38**, 4075–4084 (2004).
- Park, J., Yamashita, N., Wu, G. & Tanaka, H. Removal of pharmaceuticals and personal care products by ammonia oxidizing bacteria acclimated in a membrane bioreactor: contributions of cometabolism and endogenous respiration. *Sci. Total Environ.* **605–606**, 18–25 (2017).
- Da Silva, T. H. G., Furtado, R. X. S., Zaiat, M. & Azevedo, E. B. Tandem anaerobic-aerobic degradation of ranitidine, diclofenac, and simvastatin in domestic sewage. *Sci. Total Environ.* **721**, 137589 (2020).
- Joss, A., Andersen, H., Ternes, T., Rychle, P. R. & Siegrist, H. Removal of estrogens in municipal wastewater treatment under aerobic and anaerobic conditions: consequences for plant optimization. *Environ. Sci. Technol.* **38**, 3047–3055 (2004).
- Tisler, S. & Zwiener, C. Aerobic and anaerobic formation and biodegradation of guanil urea and other transformation products of metformin. *Water Res.* **149**, 130–135 (2019).
- Joss, A. et al. Biological degradation of pharmaceuticals in municipal wastewater treatment: proposing a classification scheme. *Water Res.* **40**, 1686–1696 (2006).
- Cooper, M. M., Elzerman, A. W. & Lee, C. M. Teaching chemistry in the new century: environmental chemistry. *J. Chem. Educ.* **78**, 1169–1169 (2001).
- Brown, A. K., Ackerman, J., Cicek, N. & Wong, C. S. In situ kinetics of human pharmaceutical conjugates and the impact of transformation, deconjugation, and sorption on persistence in wastewater batch bioreactors. *Environ. Pollut.* **265**, 114852 (2020).
- Radjenović, J., Petrović, M. & Barceló, D. Fate and distribution of pharmaceuticals in wastewater and sewage sludge of the conventional activated sludge (CAS) and advanced membrane bioreactor (MBR) treatment. *Water Res.* **43**, 831–841 (2009).
- Fan, H., Li, J., Zhang, L. & Feng, L. Contribution of sludge adsorption and biodegradation to the removal of five pharmaceuticals in a submerged membrane bioreactor. *Biochem. Eng. J.* **88**, 101–107 (2014).
- Krauss, M., Longrée, P., Dorusch, F., Ort, C. & Hollender, J. Occurrence and removal of N-nitrosamines in wastewater treatment plants. *Water Res.* **43**, 4381–4391 (2009).
- Wijekoon, K. C. et al. Removal of N-nitrosamines by an aerobic membrane bioreactor. *Bioresour. Technol.* **141**, 41–45 (2013).
- Brakstad, O. G. et al. Biotransformation in water and soil of nitrosamines and nitramines potentially generated from amine-based CO<sub>2</sub> capture technology. *Int. J. Greenh. Gas. Control* **70**, 157–163 (2018).
- Alvarino, T., Suarez, S., Lema, J. M. & Omil, F. Understanding the removal mechanisms of PPCPs and the influence of main technological parameters in anaerobic UASB and aerobic CAS reactors. *J. Hazard. Mater.* **278**, 506–513 (2014).
- Mazioti, A. A., Stasinakis, A. S., Gatidou, G., Thomaidis, N. S. & Andersen, H. R. Sorption and biodegradation of selected benzotriazoles and hydroxybenzothiazole in activated sludge and estimation of their fate during wastewater treatment. *Chemosphere* **131**, 117–123 (2015).
- Loganathan, B. G., Sajwan, K. S., Sinclair, E., Senthil Kumar, K. & Kannan, K. Perfluoroalkyl sulfonates and perfluorocarboxylates in two wastewater treatment facilities in Kentucky and Georgia. *Water Res.* **41**, 4611–4620 (2007).
- Urase, T. & Kikuta, T. Separate estimation of adsorption and degradation of pharmaceutical substances and estrogens in the activated sludge process. *Water Res.* **39**, 1289–1300 (2005).
- Abegglen, C. et al. The fate of selected micropollutants in a single-house MBR. *Water Res.* **43**, 2036–2046 (2009).
- Xue, W. et al. Elimination and fate of selected micro-organic pollutants in a full-scale anaerobic/anoxic/aerobic process combined with membrane bioreactor for municipal wastewater reclamation. *Water Res.* **44**, 5999–6010 (2010).
- Stevens-Garmon, J., Drewes, J. E., Khan, S. J., McDonald, J. A. & Dickenson, E. R. Sorption of emerging trace organic compounds onto wastewater sludge solids. *Water Res.* **45**, 3417–3426 (2011).
- Fernandez-Fontaina, E., Pinho, I., Carballa, M., Omil, F. & Lema, J. M. Biodegradation kinetic constants and sorption coefficients of micropollutants in membrane bioreactors. *Biodegradation* **24**, 165–177 (2013).
- Chakraborty, T. EC3: Combining clustering and classification for ensemble learning. *Proc. IEEE Int. Conf. Data Min. ICDM* **2017**, 781–786 (2017).
- Lu, J., Li, H., Tu, Y. & Yang, Z. Biodegradation of four selected parabens with aerobic activated sludge and their transesterification product. *Ecotoxicol. Environ. Saf.* **156**, 48–55 (2018).
- Gallé, T. et al. Large-scale determination of micropollutant elimination from municipal wastewater by passive sampling gives new insights in governing parameters and degradation patterns. *Water Res.* **160**, 380–393 (2019).
- Douziech, M. et al. Quantifying variability in removal efficiencies of chemicals in activated sludge wastewater treatment plants – a meta-analytical approach. *Environ. Sci. Process. Impacts* **20**, 171–182 (2018).
- Rich, S. L., Zumstein, M. T. & Helbling, D. E. Identifying functional groups that determine rates of micropollutant biotransformations performed by wastewater microbial communities. *Environ. Sci. Technol.* **56**, 984–994 (2022).
- Kim, G. A., Son, H. J., Kim, C. W. & Kim, S. H. Nitrosamine occurrence at Korean surface water using an analytical method based on GC/LRMS. *Environ. Monit. Assess.* **185**, 1657–1669 (2013).
- Pomiès, M., Choubert, J. M., Wisniewski, C. & Coquery, M. Modelling of micropollutant removal in biological wastewater treatments: a review. *Sci. Total Environ.* **443**, 733–748 (2013).
- Giuseppe, V. MiniSom: minimalist and NumPy-based implementation of the Self Organizing Map. <https://github.com/JustGlowing/minisom/> (2018). Accessed on 21 March 2022.
- Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

## ACKNOWLEDGEMENTS

This study was supported by the Korea Environment Industry & Technology Institute through the “Project for developing innovative drinking water and wastewater technologies,” funded by the Korea Ministry of Environment [Grant No. 2019002710010], and the National Research Foundation of Korea (NRF) grant, funded by the Korean government (MSIT) [No. 2021R1C1C2005643].

## AUTHOR CONTRIBUTIONS

The manuscript was written with the contributions of all authors. All authors have read and agreed to the published version of the manuscript. Each author's contributions are as follows: S.J.L.: Conceptualization, Methodology, Data analysis, Writing- Original draft, Co-first author J.S.: Conceptualization, Methodology, Writing- Original draft, Co-first author M.G.S.: Conceptualization, Data curation, Validation, Writing- Original draft. J.L.: Data curation, Validation, Reviewing. W.W.E.: Data curation, Validation. D.-H.L.: Data curation, Validation. E.J.: Data curation, Validation. S.H.C.: Writing- Reviewing and Editing. Y.L.: Writing- Reviewing and Editing. M.S.: Supervision, Funding acquisition, Writing- Reviewing and Editing. Co-corresponding author S.W.H.: Supervision, Resources, Funding acquisition, Writing- Reviewing and Editing, Co-corresponding author.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41545-023-00282-6>.

**Correspondence** and requests for materials should be addressed to Moon Son or Seok Won Hong.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023