water

MDPI

Article

Insightful Analysis and Prediction of SCOD Component Variation in Low-Carbon/Nitrogen-Ratio Domestic Wastewater via Machine Learning

Xuyuan Zhang [1], Yingqing Guo [1,2,*], Haoran Luo [1], Tao Liu [1] and Yijun Bao [1]

[1] School of Environmental Science and Engineering, Changzhou University, Changzhou 213164, China; s21030857037@smail.cczu.edu.cn (X.Z.); s21030857010@smail.cczu.edu.cn (H.L.); s21030857125@smail.cczu.edu.cn (T.L.); s21030857054@smail.cczu.edu.cn (Y.B.)
[2] School of Urban Construction, Changzhou University, Changzhou 213164, China
* Correspondence: water8820@163.com

Abstract: The rapid identification of the amount and characteristics of chemical oxygen demand (COD) in influent water is critical to the operation of wastewater treatment plants (WWTPs), especially for WWTPs in the face of influent water with a low carbon/nitrogen (C/N) ratio. Given that, this study carried out batch kinetic experiments for soluble chemical oxygen demand (SCOD) and nitrogen degradation for three WWTPs and established machine learning (ML) models for the accurate prediction of the variation in SCOD. The results indicate that four different kinds of components were identified via parallel factor (PARAFAC) analysis. C1 (Ex/Em = 235 nm and 275/348 nm, tryptophan-like substances/soluble microbial by-products) contributes to the majority of internal carbon sources for endogenous denitrification, whereas C4 (230 nm and 275/350 nm, tyrosine-like substances) is crucial for readily biodegradable SCOD composition according to the machine learning (ML) models. Furthermore, the gradient boosting decision tree (GBDT) algorithm achieved higher interpretability and generalizability in describing the relationship between SCOD and carbon source components, with an R^2 reaching 0.772. A Shapley additive explanations (SHAP) analysis of GBDT models further validated the above result. Undoubtedly, this study provided novel insights into utilizing ML models to predict SCOD through the measurements of the excitation–emission matrix (EEM) in specific Ex and Em positions. The results could help us to identify the degradation and transformation relationship between different kinds of carbon sources and nitrogen species in the wastewater treatment process, and thus provide a novel guidance for the optimized operation of WWTPs.

Keywords: wastewater treatment; low carbon/nitrogen ratio; carbon sources; PARAFAC; machine learning

check for updates

Citation: Zhang, X.; Guo, Y.; Luo, H.; Liu, T.; Bao, Y. Insightful Analysis and Prediction of SCOD Component Variation in Low-Carbon/Nitrogen-Ratio Domestic Wastewater via Machine Learning. *Water* **2024**, *16*, 1018. https://doi.org/10.3390/w16071018

Academic Editors: Mehdi Jamei and Masoud Karbasi

Received: 22 February 2024
Revised: 18 March 2024
Accepted: 26 March 2024
Published: 1 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Biological denitrification requires sufficient carbon sources to achieve higher nitrogen removal efficiency [1]. Especially in heterotrophic denitrification, carbon sources provide electron donors for the denitrification process [2]. However, the current domestic sewage in China faces the problem of a low carbon-to-nitrogen ratio (C/N) in influent water. In order to ensure the efficient removal of nitrogen, many studies have proposed that denitrification efficiency can be enhanced by additional dosing of sodium acetate, glucose, methanol, etc. [3–5]. However, the addition of external carbon sources requires a profound understanding of the COD components of the influent water in the local wastewater treatment plant (WWTP), as well as the kinetics of degradation and utilization processes [6,7].

Numerous studies have shown that different COD fractions can exert great influence on effluent quality [8–10]. For example, one of the most popular software, BIOWIN 6.2, classified COD components into non-biodegradable COD (NBCOD), readily biodegradable

COD (RBCOD), slowly biodegradable COD (SBCOD), etc. [11–13]. The efficiency of nitrogen removal can only be optimized by adding external RBCOD at the most appropriate time, for example, when NH_4^+-N is fully oxidized to NO_2^--N and NO_3^--N. Therefore, the rapid measurement and analysis of the proportion of different COD components in wastewater and its degradation process could provide a certain reference for the precise addition of carbon sources and the enhancement of nitrogen removal efficiency in wastewater plants. Three-dimensional excitation–emission matrix (3D-EEM) spectroscopy has recently been recognized as an effective method to qualitatively and quantitatively characterize dissolved organic matters (DOMs) in wastewater treatment processes [14–16]. Guo, L. et al. indicated that the NO_3^--N reduction process is closely related to the utilization of SCOD [17], and the soluble tyrosine-like proteins and tryptophan-like proteins were the most easily utilized carbon sources during the denitrification process according to the 3D-EEM analysis [18]. The above studies have identified biodegradable carbon sources through 3D-EEM, which in turn provides us with richer and more comprehensive information than SCOD in wastewater. Tang et al. [19] further assessed the correlation between PARAFAC components and water quality parameters such as COD, DOC, Chlorophyll a, TN, TP, and NH_4^+-N using Pearson's correlation coefficient and redundancy analysis (RDA). The results indicated that COD concentration was positively correlated with humic-like substances. Several studies [20,21] have sought to investigate the potential of fluorescence excitation–emission (F-EEM) spectroscopy as an alternative analytical method for assessing the relationship between the presence of crucial drugs of addiction, ammonia, and pH with PARAFAC components. However, the correlation analysis techniques used in the above studies were RDA, canonical correlation, or linear correlation analysis methods.

In actual fact, the intricate mechanisms of fluorescence and the presence of irregular phenomena, which pose challenges for explanation, have raised concerns among researchers regarding the adequacy of linear relationship modeling in characterizing dissolved organic matter (DOM) during the analysis of EEM data [22]. In order to accurately predict the carbon source components and concentration during the different stages of wastewater treatment, it is of vital importance to establish the correlation between SCOD and different EEM components at specific excitation and emission wavelengths. Therefore, in the face of information-rich 3D-EEM data, the application of ML would be promising in establishing a prediction model for components of EEM extracted by PARAFAC and SCOD components, which can effectively reduce the amount of prediction for water quality analysis by extracting the main analytical constituents, and thus provide a technical reference for the operation and control of wastewater plants.

Recently, machine learning (ML), which is regarded as one of the technical approaches of artificial intelligence, has demonstrated unique performance in describing the relationship between various input and output factors [23,24]. The application of ML in fields including water quality prediction [25,26], water source classification [24], source tracing, and the aerobactin assessment of contaminants [27,28] has received significant attention in previous studies. ML involves a variety of algorithmic models, such as artificial neural networks (ANNs) [29], K-nearest neighbors (kNN) [30], decision tree (DT) [31], and gradient boosting decision trees (GBDTs) [32]. These models are implemented to achieve the predictive control of effluent quality such as biological oxygen demand (BOD), chemical oxygen demand (COD), and nutrient concentration. According to previous studies [33–37], a dynamic kernel extreme learning machine was proposed, including 170 samples and eight variables, to predict the COD proportion of industrial wastewater, and achieved a 10-fold cross-validation R^2 of 0.708 [38]. Alavi et al. [39] proposed a novel computing algorithm that integrates an intelligent optimization algorithm with a KELM for the prediction of inlet COD concentrations in WWTPs. This study also compared the performance of different algorithms for optimizing real-time COD prediction. Zhao et al. [40] utilized six kinds of machine learning models to estimate the RBCOD and SBCOD in municipal wastewater with an R^2 higher than 0.80 by inputting oxidation–reduction potential (ORP) data. Kim et al. [41] established high-performance (>95% of accuracy) ML models to predict the influ-

ence of different feeding carbon sources (acetate, glucose, and starch) on the microcosm communities of activated sludge. Therefore, machine learning has demonstrated excellent performance in past research; these algorithms help us to identify the complex relationship between input factors and output factors, and achieve high levels of accuracy and generalization. Nevertheless, they were in a superficial level of ML application, that is, merely focusing on predicting and analyzing simple important features. Moreover, few research studies reported a practical technical reference to assist WWTPs to guide the operation and improve the SCOD removal performance.

To provide a detailed analysis of the degradation, transformation, and removal processes of carbon sources in three WWTPs facing a low C/N problem, we analyzed the degradation process of SCOD in aerobic and anoxic processes by conducting batch experiments in the lab, and identified the key carbon sources that could be easily degraded using PARAFAC technology. In addition, we developed a GBDT-based model related to SCOD and four PARAFAC components through comparing four ML models. We also compared the four models in terms of model interpretation using Shapley additive explanations (SHAP) analysis to avoid the problem of an ML "black box". Furthermore, to make the developed ML model more practical, we identified PARAFAC components closely related to SCOD removal in wastewater, and attempted to accurately predict the carbon source degradation process of the target WWTPs. The ML model proposed in this study is expected to be a valuable tool for further application in process optimization and accurate carbon source dosing for practical WWTPs.

2. Materials and Methods

2.1. Inoculated Sludge and Experimental Water Samples

Sludge samples are obtained from 3 sewage plants (Lanzhou, Gansu, China), which include the activated sludge process of wastewater treatment plant-A (WWTP-A), the oxidation ditch process of wastewater treatment plant-B (WWTP-B), and the anaerobic–anoxic–aerobic process (AAO) of wastewater treatment plant-C (WWTP-C). The total hydraulic retention time (HRT) of the effluent within the reaction ranged from 9 to 11 h. Subsequently, the effluent after enhanced primary treatment was fully mixed and disinfected using ultraviolet to meet the national secondary standards, and then discharged. The quality of the influent of the water treatment plants is shown in Table 1.

Table 1. Treatment scale and influent water quality of sewage plants.

WWTP	Treatment Capacity (Tons/Day)	COD (mg/L)	BOD_5 (mg/L)	NH_4^+-N (mg/L)	TN (mg/L)	TP (mg/L)
WWTP-A	160, 350 ± 138	189.3 ± 31.3	91.7 ± 12.7	32.8 ± 5.8	34.5 ± 6.1	5.7 ± 0.9
WWTP-B	160, 760 ± 157	171.2 ± 26.7	106.5 ± 15.6	28.9 ± 6.2	33.6 ± 6.9	4.3 ± 1.1
WWTP-C	40, 280 ± 39	268.5 ± 36.9	152.6 ± 18.9	42.3 ± 7.3	45.1 ± 8.7	6.1 ± 1.4

2.2. Experimental Setup and Operational Parameters

The experimental sludge samples were taken from the biochemical sections of three typical sewage plants in a certain region, which were proportionately mixed with domestic sewage in a 2 L cylindrical reactor. In addition, the MLSS of sludge was 4–5 g/L, the aerobic aeration time was determined to be 480 min, which was consistent with the residence time in the aerobic section of the sewage plants, while the anaerobic treatment time was 260 min, and the reaction temperature was controlled at 25 °C. The kinetic experiment was conducted in reactor-A (R-A), reactor-B (R-B), and reactor-C (R-C) for WWTP-A, WWTP-B, and WWTP-C, respectively. The water samples were taken at different times during aerobic and anoxic processes, and the experiment was repeated four times for each treatment process, which were denoted as S1, S2, S3, and S4, respectively.

2.3. Water Quality Measurement and Parallel Factor Analysis

The partial samples were filtered using a 0.45 μm membrane before being digested with DRB200 digestor (Hach, Loveland, USA); the Hach-DR6000 water quality analyzer (Hach, Loveland, CO, USA) was then utilized to collect the SCOD values. The NH_4^+-N, NO_2^--N, NO_3^--N, and MLSS were measured according to standard methods (APHA, 2005) in this study [42]. The three-dimensional fluorescence spectra of the samples were measured using an F7100 fluorescence spectrophotometer (Hitachi, Toyko, Japan) before treating them with a 0.45 μm membrane. The PTM voltage was set at 600 V; the emission (Em) and excitation (Ex) wavelengths were scanned in the ranges of 200–550 nm and 200–500 nm, respectively, with the scanning step of 5 nm; and the scanning speed was controlled at 12,000 nm/min. Furthermore, all three-dimensional fluorescence data were differenced and subtracted from ultrapure water data prior to analysis to deduct Raman scattering. Then, the parallel factor analysis (PARAFAC) was employed using MATLAB 7.1 software embedded with DOMFluor toolbox to identify the components of SCOD [43].

2.4. Machine Learning Model

The redundancy analysis (RDA) of different fluorescence components and SCOD was firstly performed using CANOCO 5.0 software in order to pre-analyze their correlation. Four kinds of prevalent machine learning algorithms, including support vector machine (SVM), decision tree (DT), random forest (RF), and gradient boosting decision tree (GBDT), were exploited for modeling, followed by an assessment of model applicability by comparing fitting accuracy. Finally, the optimal model was selected for focused analysis [44]. Among them, the SVM algorithm was employed to establish a particular optimal decision hyperplane and maximize the 2 closest classes on both sides of this plane and itself to generalize each classification [45]. Meanwhile, the other three algorithms, based on Classification and Regression Trees (CARTs), completed the prediction by constructing a binary tree to recursively divide the samples, and the split node was established according to the minimum variance of the samples. All the above algorithms are described in detail in the literature [46,47].

To prevent overfitting in the machine learning modeling process, a ten-fold cross-validation (10-fold CV) method was used to adapt the hyperparameters, where the training data were divided into 10 subsets, and during each training iteration, one subset was used as the validation set, and the others constituted the training set to improve the prediction and generalization ability of the model [48]. In addition, the model performance was evaluated using the coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE), as calculated in Equations (1)–(3). The higher the R^2 and the lower the MSE and MAE, the higher the model accuracy.

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(\hat{y} - y)^2}{\sum_{n=1}^{N}(\hat{y} - \overline{y})^2} \quad (1)$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2 \quad (2)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | \hat{y} - y | \quad (3)$$

where $\hat{y}$, y, and $\overline{y}$ are the predicted, actual, and average values of the target feature, respectively, n is the number of data points for any particular instance, and N is the total number of data points.

2.5. SHAP Analysis

Shapley additive explanations (SHAP) analysis was performed to enhance the reliability and validity of the results calculated by the ML algorithm [49]. Therefore, the SHAP value, determined by the optimum Shapley value from game theory, offers a rational way to allocate payoffs among coalition members [50]. A subset of the eigenvalues and predictions

of the subset represent the coalition and payoffs of the coalition, respectively. The SHAP value is calculated as follows:

$$g(z') = \varphi_0 + \sum_{i=1}^{N} \varphi_i z_i' \quad (4)$$

where N indicates the maximum subset size of the feature values, φ_i represents feature i's Shapley value, and $z'\epsilon$, z_i' relate to the simplified features, with 0 or 1 indicating whether the feature is absent or present during the Shapley value calculation.

3. Discussion

3.1. SCOD Removal Performance

Figure 1 shows the concentration and removal of SCOD during aerobic and anoxic processes. As expected, SCOD was gradually degraded during aerobic treatment, and the lowest SCOD concentrations were observed at 480 min in the range of 8–30, 13–24, and 30–41 mg/L, respectively. A rapid degradation of SCOD was observed during the early 100 min and then the degradation gradually stabilized from 100 to 480 min. The treated wastewater was subsequently transferred to an anaerobic stage for denitrification after 480 min. After 260 min of anaerobic treatment, the removal of SCOD was stable, with removal efficiencies of 60.9–84.5%, 66.7–81.5%, and 70.8–83.8%, respectively.

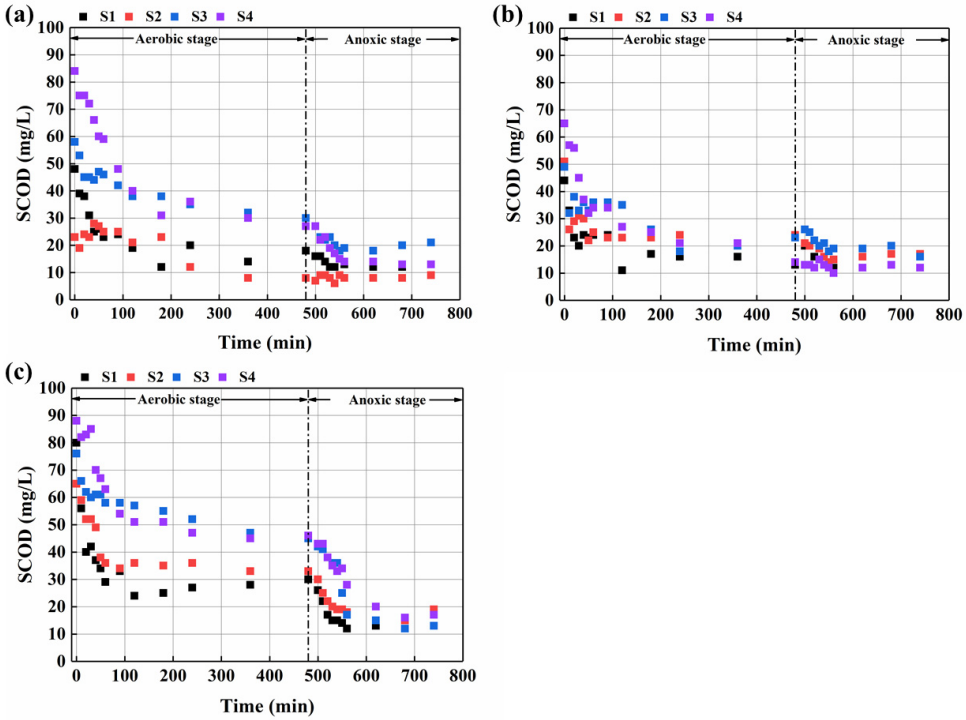


Figure 1. The kinetic experiment of SCOD during the aerobic and anoxic processes of (**a**) R-A, (**b**) R-B, and (**c**) R-C.

Compared to R-A and R-B, R-C exhibited a relatively poor capacity of SCOD removal, with 37.5–62.1% during the aerobic stage. However, SCOD removal during the anaerobic stage was significantly improved in R-C. The discrepancy between SCOD removal in different reactors can be attributed to the differences in dissolved organic matter (DOM) composition between wastewater samples. It was reported that 78.1–86.5% of total COD and 82.6–86.6% of total TOC consist of DOM in the effluents of WWTPs [51]. The aerobic degradation capacity of the SCOD of microbial communities in different biological treatment processes varies with the composition of DOM. Generally, according to EEM spectra analysis, the protein-like fractions were easily degraded during biological treatment, but fulvic-like and humic-like factions were more biologically recalcitrant [52–54].

Moreover, the intensities of fulvic-like and humic-like fluorescence was even increased after the bio-treatment [54].

Figure 2 displays the performance of nitrogen removal. In the aerobic stage, the final NH_4^+-N removal efficiencies were 97.0%, 89.1%, and 91.3%, respectively. The total inorganic nitrogen concentrations of the effluent were 12.8, 11.4 and 16.6 mg/L, resulting in average nitrogen removal efficiencies of 18.0%, 32.8%, and 37.3% for R-A, R-B, and R-C, respectively. Despite the lower nitrification efficiency of R-C, it did really achieve higher denitrification efficiency due to the higher nitrogen concentration in the influent compared with R-A and R-B. On the one hand, the influent TN of WWTP-C was 43.2 mg/L, which was higher than that of WWTP-A (33.9 mg/L) and WWTP-B (32.2 mg/L). On the other hand, the average BOD_5/COD (B/C) ratios of the influent for WWTP-A, WWTP-B, and WWTP-C were 0.48, 0.62, and 0.57, respectively. The higher B/C reflects the poorer biodegradability of the influent for WWTP-C, even with a higher BOD_5. Considering the absence of an additional carbon source during the anoxic stage, it can be speculated that the remaining SCOD in WWTP-C was more susceptible to utilization by microorganisms as an internal carbon source, contributing to endogenous denitrification.

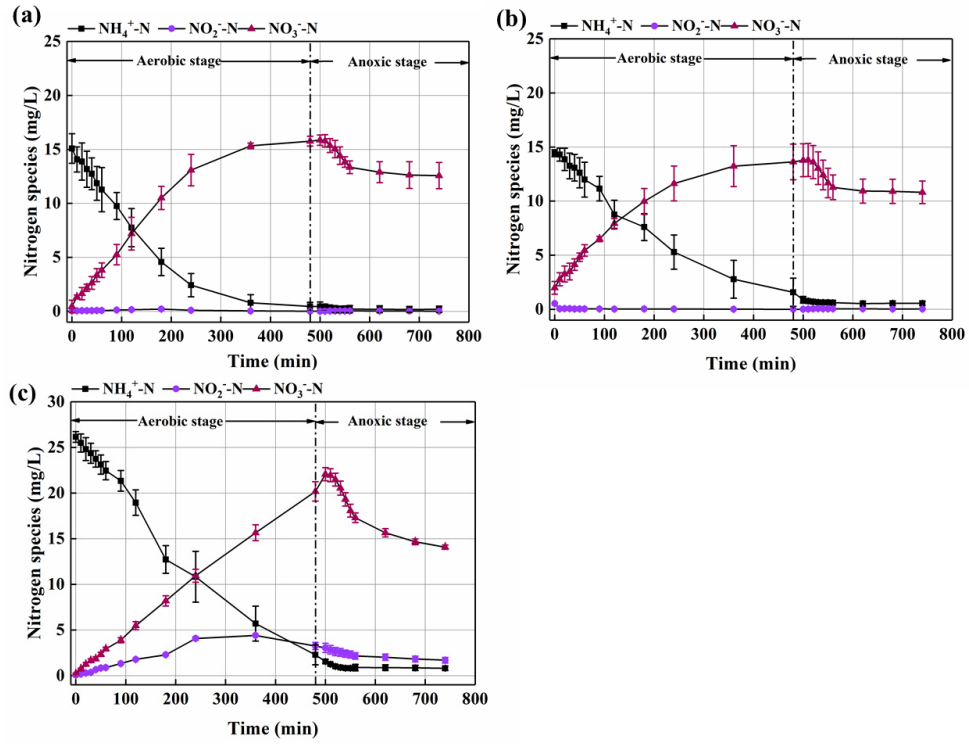


Figure 2. The variation in NH_4^+-N, NO_3^--N, and NO_2^--N during the aerobic and anoxic processes in (**a**) R-A, (**b**) R-B, and (**c**) R-C. (The length of the vertical line at each point represents the standard deviation).

3.2. Degradation of Different Fluorescent Components in Reactor Effluent

3.2.1. Component of DOM Identified through PARAFAC Analysis

The PARAFAC analysis was used to reveal the DOM composition change during aerobic and anoxic biological treatment due to its remarkable DOM characterization ability. A total of 276 samples were inputted to DOMFluor toolbox in Matlab7.1 and decomposed into four fluorescent components (Table 2). The corresponding contour plots and the excitation (Ex) and emission (Em) loadings are presented in Figure 3. Component 1 (C1) showed two excitation peaks at 235 and 275 nm with one emission peak at 348 nm, which can be assigned to tryptophan-like substances/soluble microbial by-products [55,56]. Component 2 (C2) displayed two excitation maxima (245 and 315 nm) corresponding to the same emission maximum at 400 nm. The peaks of C2 are related to terrestrial humic acid-like

substances (terrestrial HA) [57,58]. Component 3 (C3) had excitation and emission at (245, 315)/450 nm, similar to microbial fulvic acid-like substances (microbial FA) [59,60]. The peaks of C2 (Ex/Em = (230, 275)/350 nm), similar to the peaks of the C1, were characterized as tyrosine-like substances [61].

Table 2. Component results of effluent samples.

Component	Excitation (nm)	Emission (nm)	Compound
C1	235/275	348	tryptophan-like substances/soluble microbial by-products
C2	245/315	400	terrestrial HA
C3	275/360	450	microbial FA
C4	230/275	350	tyrosine-like substances

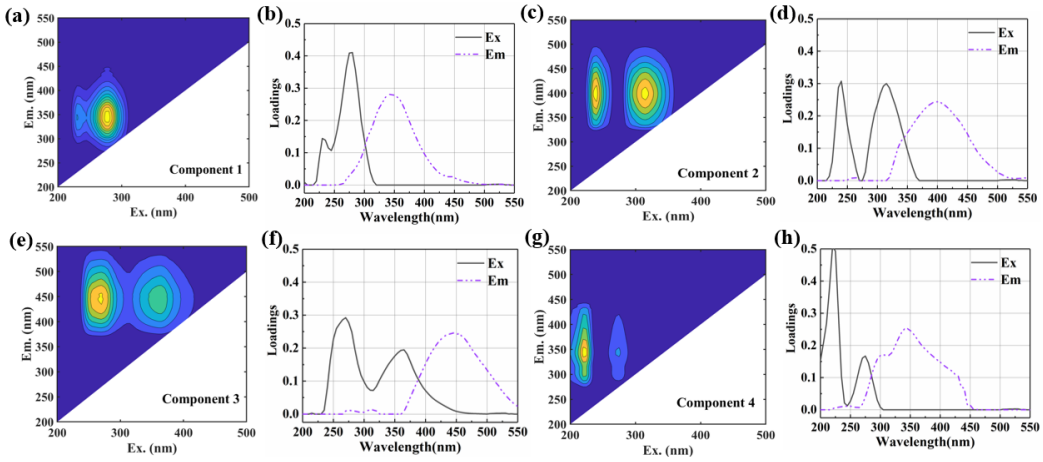


Figure 3. EEM contours and loadings of four components identified through the PARAFAC of effluent samples in three treatment reactors. (**a**,**b**) for C1; (**c**,**d**) for C2; (**e**,**f**) for C3; and (**g**,**h**) for C4.

3.2.2. DOM Fractions in Influent and Effluent Samples

In order to investigate the change patterns of different internal carbon sources during the degradation process, the degradation dynamics of F_{max} of four individual fluorescent components originating from PARAFAC were analyzed. Figure 4 shows the F_{max} values of the PARAFAC-derived components of influent and effluent samples. Components C1 and C4 significantly contributed to fluorescent intensity, which accounted for 33.4–42.0% and 29.7–40.3% of total F_{max} values for the influent sample, respectively. This revealed that tryptophan-like and tyrosine-like proteins dominated the constituents of influent samples. The result is in agreement with the EEM results reported in a previous study about domestic wastewater [62]. It is worth noting that the fluorescent intensities of components C1 and C4 in the influent of WWTP-C were generally higher than those of the other two WWTPs. This is the main factor that contributed to the poor removal ability of SCOD for WWTP-C. With respect to components C2 and C3, they accounted for 14.8–15.8% and 11.6–12.5% of total F_{max} values for the influent sample, respectively. It was reported that the humic-like components C2 and C3 were characterized by a high aromatic nature and unsaturation [63]. Except for component C3, the degradation efficiencies of the fluorescent components follow the order of C4 > C1 > C2, which reached 93.9~100%, 19.4~43.1%, and 12.3~18.7%, respectively. The F_{max} of component C3 even exhibited an increasing trend after the aeration process, which indicated that component C3 may be generated from the metabolism of microorganisms [63,64].

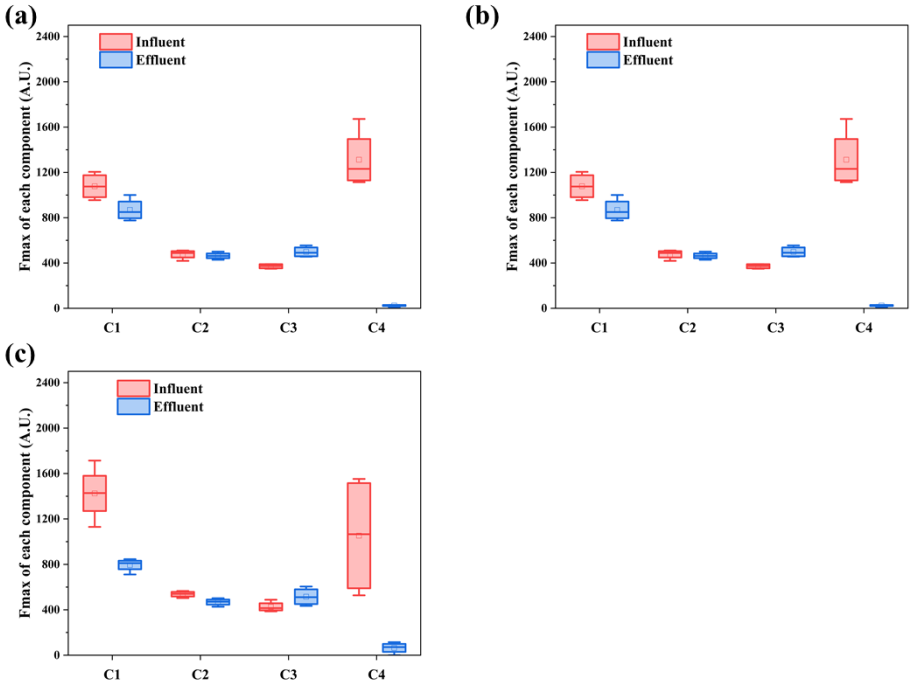


Figure 4. The F_{max} of C1–C4 fluorescent components in influent and effluent samples of (**a**) R-A, (**b**) R-B, and (**c**) R-C.

3.2.3. Degradation Kinetics Analysis of PARAFAC Components

The changes in different fluorescent components during the aerobic/anoxic process are shown in Figure 5. The SCOD concentrations in all reactors showed decreasing trends during the aerobic/anoxic process according to Figure 1; however, the DOM components with different fluorescent characteristics varied. The intensity of C4 significantly decreased with prolonged aeration, suggesting the effectiveness of the aeration procedure for C4 degradation. This phenomenon is consistent with changes in SCOD (Figure 1). It has been reported that C4 (tyrosine-like substance) was concentrated in the low molecular weight fraction [65]. This property contributed to its great bioavailability during aeration [66]. On the contrary, C2 and C3, representing humic-like and fulvic-like substances, were generally recalcitrant to biological treatment. Both C2 and C3 were ineffectively removed in either aerobic or anoxic stages. The F_{max} value of C2 increased during the first 40 min and gradually returned to the same level as the initial value. Moreover, the F_{max} values of C3 increased during the aeration treatment and reached 34.4%, 21.4%, and 39.1%, which were higher than those of the influent. This was consistent with a previous study that described the microbial transformation of protein-like substances in SMP into fulvic-like and humic-like substances, which might contribute to the low and negative removal of fulvic-like and humic-like substances [67].

As for C1, the F_{max} value decreased by 19.4%, 46.7%, and 27.7% after aeration, respectively. The initial increasing trends of R-A and R-C could be attributed to the degradation and diversification of proteins with higher molecular weight by microbes, resulting in the production of lignin, polyphenol analogs, and nitrogen-containing molecules [68]. The variation of C1 (i.e., protein-like DOM) removal between three groups might be affected by the type of influent [68]. Notably, C1 significantly decreased in R-C during the anoxic stage, which was compatible with SCOD removal. This indicates that C1 is closely related to the internal carbon source for endogenous denitrification. Moreover, the initial increase and subsequent degradation of C1 in R-C indicates that the biological treatment process promoted the transformation and generation of new C1 while simultaneously decomposing it. The consumption of the hydrolysis product of the slowly biodegradable COD substrate gradually dominated its generation during the aerobic treatment. Ding et al. [69] also

found that protein-like substances (Ex/Em of 275/350 nm) significantly changed in the denitrification process, whereas humic-like and fulvic acid-like substances for component 2 remained relatively stable. The above results further indicated that protein-like substances were carbon sources which can provide easily accepted electrons for denitrifying bacteria [70]. Similar observations were made by previous studies [71,72]. This was reported to be beneficial to the growth of zoogloeal, as well as the supplementation of insufficient influent carbon sources during the subsequent denitrification [73]. Overall, the effective utilization of such slow biodegradable organic matter can make up for the lack of carbon sources in the influent and reduce the consumption of carbon source addition in the anoxic stage.

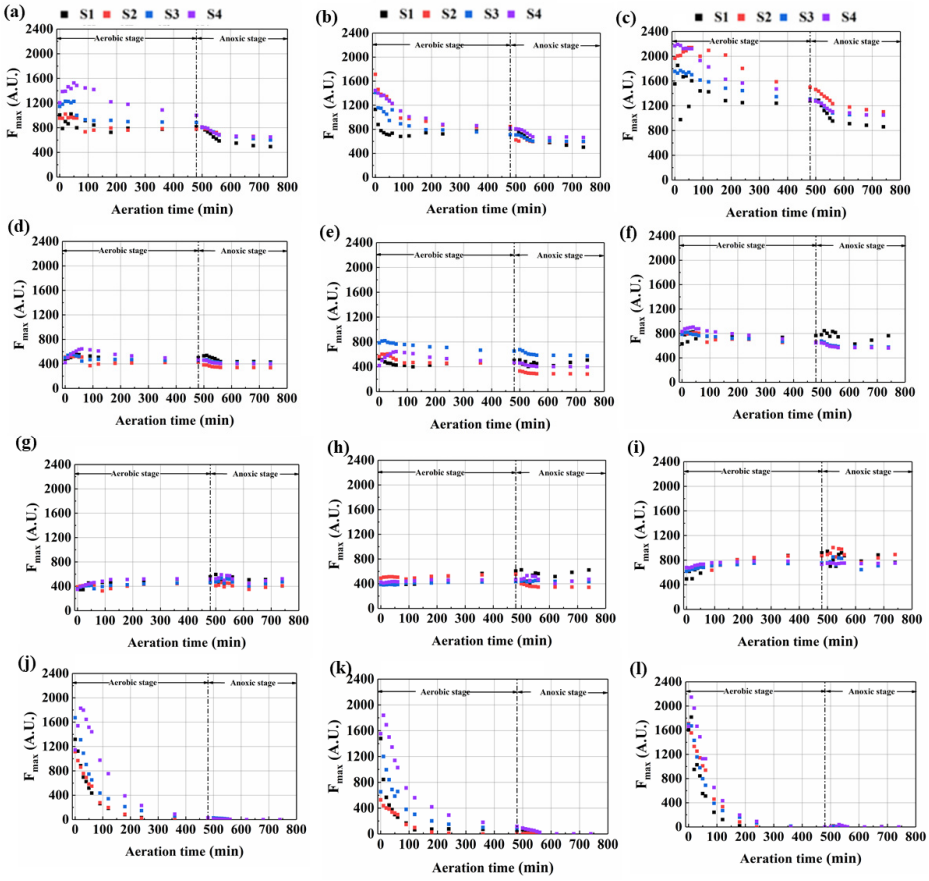


Figure 5. The F_{max} value variation in PARAFAC components of different groups. ((**a**,**d**,**g**,**j**) for C1,C2, C3, and C4 in R-A; (**b**,**e**,**h**,**k**) for C1, C2, C3, and C4 in R-B, and (**c**,**f**,**i**,**l**) for C1, C2, C3, and C4 in R-C).

3.3. Redundancy Analysis

RDA was employed to quantify the relationship between different indices and further determine the input variable for ML model establishment. Figure 6 shows the relative significance of each fluorescent component to SCOD and the nitrogen species. The contributions of RDA1 and RDA2 were 53.05% and 15.54%, respectively [74]. The angles between SCOD and other indices reveal that the four fluorescent components had positive correlations with SCOD concentration, but NO_3^--N was negatively correlated with SCOD. The length of the arrow indicates that components C1, C2, and C4 made greater contribution to SCOD than C3 [75]. This result demonstrates that C3 was most likely the hydrolysis product of SCOD. For the nitrogen species, C1, C2, and C4 showed a negative correlation with NO_3^--N, which suggested that C1, C2, and C4 may contribute to the denitrification process. C4, in particular, was negatively correlated with both NO_3^--N and NO_2^--N, which indicates that C4 played a significant role in nitrogen removal. Therefore, the indices of C1, C2, and C4 were selected as the input variables of the machine learning models. Moreover, both

C1 and C4 were negatively correlated with the aeration time. This sheds light on the fact that the aeration time could be optimized to abate C1 degradation based on the effective degradation of C4.

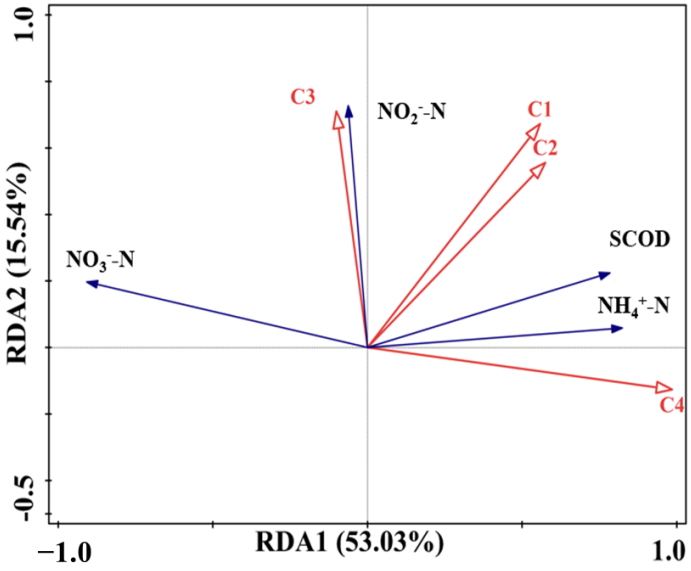


Figure 6. RDA analysis between various indices.

3.4. ML Models for SCOD Degradation Prediction Based on Fluorescent Components

The relationship between SCOD and different EEM components can be distinguished using RDA analysis; however, there exists a limitation in describing the detailed quantitative relationship between SCOD, PARAFAC components, and nitrogen transformation. For further application, ML was used to establish models for SCOD prediction during the biological treatment. The database contained 276 sets of data, and four ML models, including SVM, DT, RF, and GBDT, were employed to compare the forecasting effectiveness of SCOD. The scatter matrixes of the measured and predicted SCOD of different ML models are depicted in Figure 7. According to Figure 7, the performances of these models were reflected by R^2 conforming to GBDT > RF > DT > SVM. In detail, GBDT outperformed RF, DT, and SVM (4.30–38.0% higher R^2, 10.47–94.05 lower MSE, and 0.38-2.87 lower MAE) in test groups, as the other three models were unable to learn in a time-dependent manner and had weaker nonlinear mapping capabilities [76]. As an algorithm based on boosting ensemble, the GBDT model has the ability to reduce bias errors commonly found in conventional machine learning models [77].

3.5. Sensitivity Analysis of Fluorescent Components

Figure 8 is a SHAP summary plot which demonstrates the distribution of the SHAP values for component features and indicates the corresponding influences. With regard to the ranking of feature importance, the contributions of fluorescent components in different ML models are always C4, C1, and C2, in descending order. Therefore, the C4 of EEM spectra is crucial for SCOD composition. This further demonstrates that tyrosine-like substances are the major organic matters of SCOD, and component C4 is the key component for nitrogen removal. From Figure 8, an increase in the predictive value of F_{max} is associated with an increase in C4 and C1, which is consistent with previous studies. The high F_{max} values of C3 and C4 positively correlate with SCOD, especially the high F_{max} values of component C4, which shows long right tails in the summary plots [78]. However, the four models present different influence trends between C2 and F_{max}. Based on the SVM model, the predictive F_{max} value increases with a decrease in C2. In contrast, based on the other three models, the predictive value of F_{max} is associated with an increase in C2. According to a previous study, humic-like substances have been identified as unavailable carbon

sources that critically inhibit the biodegradability of aerobic treatment [79]. Therefore, the GBDT-based model developed in this study presents more reasonable trends than the other three models. Furthermore, the range of SHAP values in the GBDT-based SHAP plot is much larger than that in the SVM-based and DT-based SHAP plot. Previous studies [80,81] also acquired further explanations to avoid the disadvantages of the "black box" of ML models. This may be one of the reasons why the GBDT-based model is more accurate that the other models in this study.

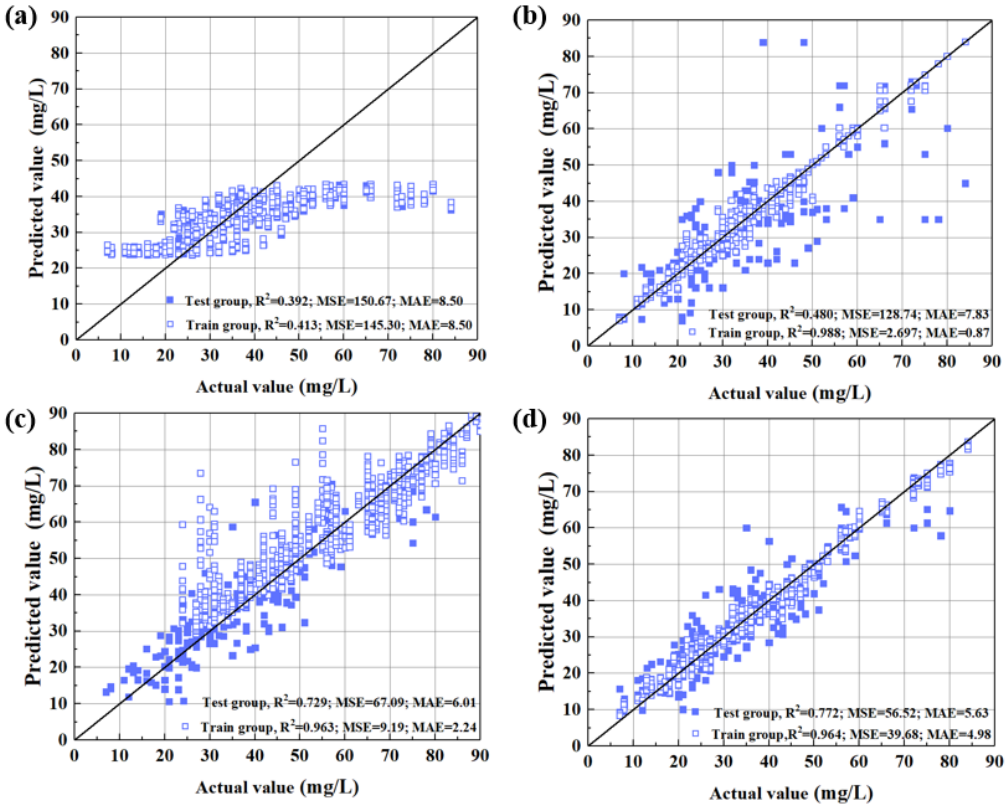


Figure 7. SCOD prediction by (**a**) SVM, (**b**) DT, (**c**) RF, and (**d**) GBDT models.

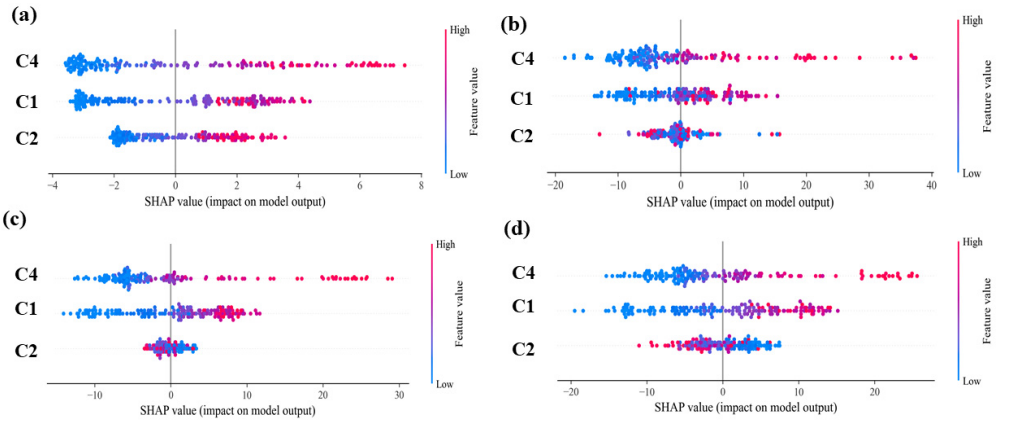


Figure 8. The SHAP summaries for the (**a**) SVM, (**b**) DT, (**c**) RF, and (**d**) GBDT models. We set the same range of x-axis values for each data set, which is more convenient for comparing the range of Shapley values.

The combination of ML models and PARAFAC analysis of EEM spectra revealed the specific fluorescent components, which demonstrated considerable importance. Based on the GBDT model, this study constructed robust mapping relationships between SCOD and

the C4 of EEM spectra. Figure 9 illustrates the correlation relationship between SCOD and the F_{max} of C4. The notable accuracy of the established models can be observed. The values of R^2 can reach 0.768, 0.725, and 0.665 for R-A, R-B, and R-C, respectively. Undoubtedly, the C4 of the internal carbon source can be an alternative input indicator to predict the degradation of SCOD and nitrogen removal combined with the GBDT model. Furthermore, the regular monitoring of the influent EEM spectra can facilitate the optimization of the process operation.

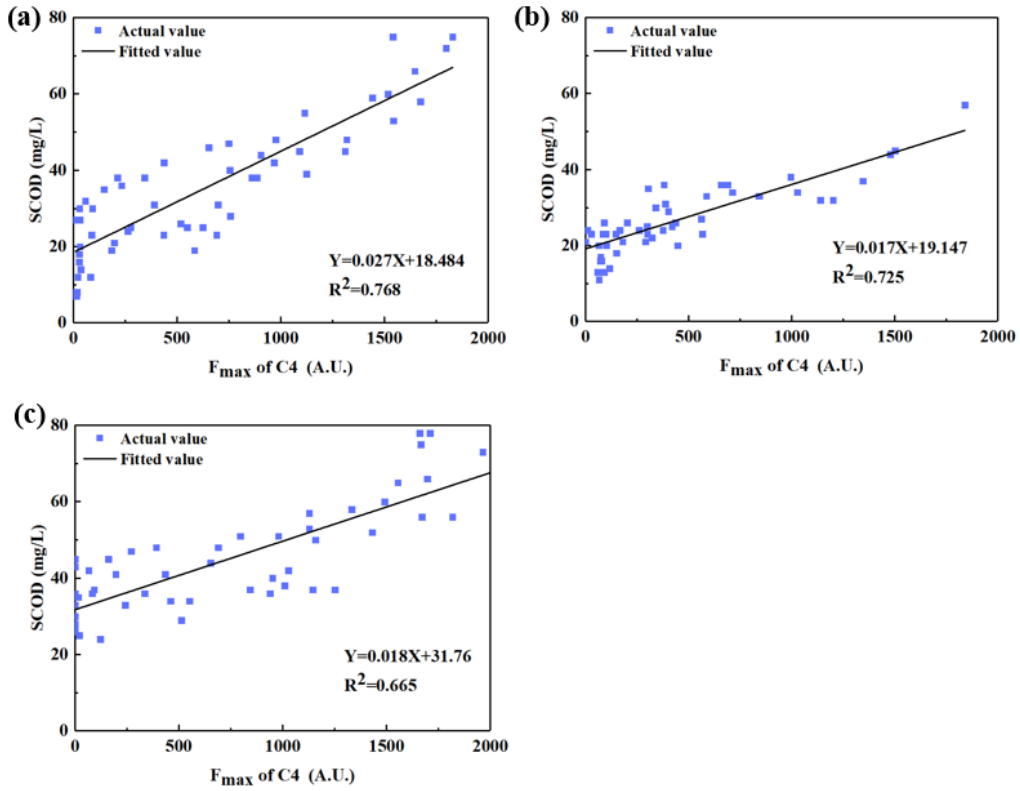


Figure 9. The linear fitting of SCOD and F_{max} value of C4 ((**a**) for R-A, (**b**) for R-B, and (**c**) for R-C).

4. Conclusions

(1) Four different kinds of DOM components were identified via 3DEEM-PARAFAC. Among them, C1 (Ex/Em = 235 and 275/348, tryptophan-like substances/soluble microbial by-products) acts as an electron donor for endogenous denitrification in the anaerobic stage, whereas C4 (230 and 275/350, tyrosine-like substances) is crucial for SCOD composition with rapid biodegradability in the aerobic stage.

(2) ML algorithms were successfully used to describe the relationship between SCOD and PARAFAC components, and the GBDT algorithm exhibited superior prediction performance with a higher R^2 of 0.772. The SHAP analysis further explained the model results and confirmed the key components (C4 > C1 > C2) related to SCOD.

(3) For the corresponding WWTPs studied in this research, the F_{max} of C1 and C4 can be selected as indicator parameters for the concentration levels of carbon sources for denitrification and regarded as a reference for the dosage and timing point of external carbon sources.

(4) As for the future development of this research, it is urgent to improve the prediction accuracy and generalization ability of the established models, so that a better matching degree of this model to the wastewater of different WWTPs can be achieved.

Author Contributions: Conceptualization, methodology, and software, X.Z.; validation, Y.G. and X.Z.; resources, data curation, writing—review and editing, X.Z.; visualization, H.L.; supervision, Y.G. and Y.B.; project administration, T.L. All authors have read and agreed to the published version of the manuscript.

Funding: Gansu Provincial Science and Technology Program Project Natural Science: 20JR10RA441.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: Thanks to the Matlab7.1 software developers at MathWorks, Inc. for facilitating the PARAFAC in this research, and to the developers of the DOMFluor toolbox.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, Z.; Dai, X.; Chai, X. Effect of different carbon sources on denitrification performance, microbial community structure and denitrification genes. *Sci. Total Environ.* **2018**, *634*, 195–204. [CrossRef] [PubMed]
2. Fu, X.; Hou, R.; Yang, P.; Qian, S.; Feng, Z.; Chen, Z.; Wang, F.; Yuan, R.; Chen, H.; Zhou, B. Application of external carbon source in heterotrophic denitrification of domestic sewage: A review. *Sci. Total Environ.* **2022**, *817*, 153061. [CrossRef] [PubMed]
3. Zhang, F.; Ma, C.; Huang, X.; Liu, J.; Lu, L.; Peng, K.; Li, S. Research progress in solid carbon source-based denitrification technologies for different target water bodies. *Sci. Total Environ.* **2021**, *782*, 146669. [CrossRef]
4. Wang, H.; Chen, N.; Feng, C.; Deng, Y. Insights into heterotrophic denitrification diversity in wastewater treatment systems: Progress and future prospects based on different carbon sources. *Sci. Total Environ.* **2021**, *780*, 146521. [CrossRef] [PubMed]
5. Cao, Q.; Li, X.; Jiang, H.; Wu, H.; Xie, Z.; Zhang, X.; Li, N.; Huang, X.; Li, Z.; Liu, X.; et al. Ammonia removal through combined methane oxidation and nitrification-denitrification and the interactions among functional microorganisms. *Water Res.* **2021**, *188*, 116555. [CrossRef] [PubMed]
6. Chen, M.; Tang, Q.; Zou, J.; Lv, X.; Deng, Y.; Ma, X.; Ma, S. Sugarcane Bagasse as Carbon Source and Filler to Enhance the Treatment of Low C/N Wastewater by Aerobic Denitrification Flora. *Water* **2022**, *14*, 3355. [CrossRef]
7. Yao, J.; Lv, S.; Wang, Z.; Hu, L.; Chen, J. Variation of current density with time as a novel method for efficient electrochemical treatment of real dyeing wastewater with energy savings. *Environ. Sci. Pollut. Res.* **2022**, *29*, 49976–49984. [CrossRef] [PubMed]
8. Zhu, Z.; Zhao, Y.; Guo, Y.; Zhang, R.; Pan, Y.; Zhou, T. A novel additional carbon source derived from rotten fruits: Application for the denitrification from mature landfill leachate and evaluation the economic benefits. *Bioresour. Technol.* **2021**, *334*, 125244. [CrossRef] [PubMed]
9. Cheng, Q.; Liu, Z.; Huang, Y.; Feng, S.; Du, E.; Peng, M.; Zhang, J. Advanced nitrogen removal performance and microbial community structure of a lab-scale denitrifying filter with in-situ formation of biogenic manganese oxides. *J. Environ. Manag.* **2023**, *331*, 117299. [CrossRef]
10. Zhang, J.; Fan, C.; Zhao, M.; Wang, Z.; Jiang, S.; Jin, Z.; Bei, K.; Zheng, X.; Wu, S.; Lin, P.; et al. A comprehensive review on mixotrophic denitrification processes for biological nitrogen removal. *Chemosphere* **2023**, *313*, 137474. [CrossRef]
11. Kim, H.; Kim, B.; Yu, J. Power generation response to readily biodegradable COD in single-chamber microbial fuel cells. *Bioresour. Technol.* **2015**, *186*, 136–140. [CrossRef] [PubMed]
12. Kilic, B.; Cecen, F. Review of experimental biodegradation data on pharmaceuticals and comparison with predictive BIOWIN models. *J. Environ. Manag.* **2023**, *344*, 118310. [CrossRef]
13. Vitanza, R.; Colussi, I.; Cortesi, A.; Gallo, V. Implementing a respirometry-based model into BioWin software to simulate wastewater treatment plant operations. *J. Water Process Eng.* **2016**, *9*, 267–275. [CrossRef]
14. Zhang, X.; Guo, P.; Yang, X.; Yao, X.; Cong, H.; Xu, B. Research on enhanced effects and mechanisms of nitrogen removal with plant carbons sources in constructed wetlands. *J. Environ. Chem. Eng.* **2023**, *11*, 110397. [CrossRef]
15. Shi, J.; Su, J.; Ali, A.; Chen, C.; Xu, L.; Yan, H.; Su, L.; Qi, Z. Nitrate removal under low carbon to nitrogen ratio by modified corn straw bioreactor: Optimization and possible mechanism. *Environ. Technol.* **2023**, *44*, 2889–2899. [CrossRef] [PubMed]
16. Liu, H.; Ouyang, F.; Chen, Z.; Chen, Z.; Lichtfouse, E. Weak electricity stimulates biological nitrate removal of wastewater: Hypothesis and first evidences. *Sci. Total Environ.* **2021**, *757*, 143764. [CrossRef] [PubMed]
17. Guo, L.; Guo, Y.; Sun, M.; Gao, M.; Zhao, Y.; She, Z. Enhancing denitrification with waste sludge carbon source: The substrate metabolism process and mechanisms. *Environ. Sci. Pollut. Res.* **2018**, *25*, 13079–13092. [CrossRef]
18. Guo, Y.; Guo, L.; Sun, M.; Zhao, Y.; Gao, M.; She, Z. Effects of hydraulic retention time (HRT) on denitrification using waste activated sludge thermal hydrolysis liquid and acidogenic liquid as carbon sources. *Bioresour. Technol.* **2017**, *224*, 147–156. [CrossRef] [PubMed]
19. Tang, J.; Li, X.; Cao, C.; Lin, M.; Qiu, Q.; Xu, Y.; Ren, Y. Compositional variety of dissolved organic matter and its correlation with water quality in peri-urban and urban river watersheds. *Ecol. Indic.* **2019**, *104*, 459–469. [CrossRef]
20. Yadav, M.K.; Aryal, R.; Short, M.D.; Saint, C.P. Fluorescence Excitation-Emission Spectroscopy: An Analytical Technique to Monitor Drugs of Addiction in Wastewater. *Water* **2019**, *11*, 377. [CrossRef]

21. Song, L.; Song, Y.; Li, D.; Liu, R.; Niu, Q. The auto fluorescence characteristics, specific activity, and microbial community structure in batch tests of mono-chicken manure digestion. *Waste Manag.* **2019**, *83*, 57–67. [CrossRef]
22. Li, L.; Wang, Y.; Zhang, W.; Yu, S.; Wang, X.; Gao, N. New advances in fluorescence excitation-emission matrix spectroscopy for the characterization of dissolved organic matter in drinking water treatment: A review. *Chem. Eng. J.* **2020**, *381*, 122676. [CrossRef]
23. Lashkaripour, A.; Rodriguez, C.; Mehdipour, N.; Mardian, R.; McIntyre, D.; Ortiz, L.; Campbell, J.; Densmore, D. Machine learning enables design automation of microfluidic flow-focusing droplet generation. *Nat. Commun.* **2021**, *12*, 25. [CrossRef]
24. Huang, R.; Ma, C.; Ma, J.; Huangfu, X.; He, Q. Machine learning in natural and engineered water systems. *Water Res.* **2021**, *205*, 117666. [CrossRef]
25. Sathasivan, A.; Kastl, G.; Korotta-Gamage, S.; Gunasekera, V. Trihalomethane species model for drinking water supply systems. *Water Res.* **2020**, *184*, 116189. [CrossRef]
26. Xia, R.; Wang, G.; Zhang, Y.; Yang, P.; Yang, Z.; Ding, S.; Jia, X.; Yang, C.; Liu, C.; Ma, S.; et al. River algal blooms are well predicted by antecedent environmental conditions. *Water Res.* **2020**, *185*, 116221. [CrossRef]
27. Balleste, E.; Belanche-Munoz, L.A.; Farnleitner, A.H.; Linke, R.; Sommer, R.; Santos, R.; Monteiro, S.; Maunula, L.; Oristo, S.; Tiehm, A.A.; et al. Improving the identification of the source of faecal pollution in water using a modelling approach: From multi-source to aged and diluted samples. *Water Res.* **2020**, *171*, 115392. [CrossRef]
28. Podgorski, J.; Berg, M. Global threat of arsenic in groundwater. *Science* **2020**, *368*, 845–850. [CrossRef]
29. Abdollahi, Y.; Zakaria, A.; Sairi, N.A.; Matori, K.A.; Masoumi, H.R.F.; Sadrolhosseini, A.R.; Jahangirian, H. Artificial neural network modelling of photodegradation in suspension of manganese doped zinc oxide nanoparticles under visible-light irradiation. *Sci. World J.* **2014**, *2014*, 726101. [CrossRef]
30. Xu, T.; Coco, G.; Neale, M. A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. *Water Res.* **2020**, *177*, 115788. [CrossRef]
31. Mori, N.; Debeljak, B.; Skerjanec, M.; Simcic, T.; Kanduc, T.; Brancelj, A. Modelling the effects of multiple stressors on respiration and microbial biomass in the hyporheic zone using decision trees. *Water Res.* **2019**, *149*, 9–20. [CrossRef]
32. Shi, S.; Xu, G. Novel performance prediction model of a biofilm system treating domestic wastewater based on stacked denoising auto-encoders deep learning network. *Chem. Eng. J.* **2018**, *347*, 280–290. [CrossRef]
33. Wei, A.; Li, D.; Bai, X.; Wang, R.; Fu, X.; Yu, J. Application of machine learning to groundwater spring potential mapping using averaging, bagging, and boosting techniques. *Water Supply* **2022**, *22*, 6882–6894. [CrossRef]
34. Huang, P.; Wang, L.; Hou, D.; Lin, W.; Yu, J.; Zhang, G.; Zhang, H. A feature extraction method based on the entropy-minimal description length principle and GBDT for common surface water pollution identification. *J. Hydroinform.* **2021**, *23*, 1050–1065. [CrossRef]
35. Abdi, J.; Hadipoor, M.; Hadavimoghaddam, F.; Hemmati-Sarapardeh, A. Estimation of tetracycline antibiotic photodegradation from wastewater by heterogeneous metal-organic frameworks photocatalysts. *Chemosphere* **2022**, *287*, 132135. [CrossRef]
36. Li, J.; Pan, L.; Suvarna, M.; Tong, Y.W.; Wang, X. Fuel properties of hydrochar and pyrochar: Prediction and exploration with machine learning. *Appl. Energy* **2020**, *269*, 115166. [CrossRef]
37. Ji, S.; Wang, X.; Zhao, W.; Guo, D. An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise. *Math. Probl. Eng.* **2019**, *2019*, 8503252. [CrossRef]
38. Liu, H.; Zhang, Y.; Zhang, H. Prediction of effluent quality in papermaking wastewater treatment processes using dynamic kernel-based extreme learning machine. *Process Biochem.* **2020**, *97*, 72–79. [CrossRef]
39. Alavi, J.; Ewees, A.A.; Ansari, S.; Shahid, S.; Yaseen, Z.M. A new insight for real-time wastewater quality prediction using hybridized kernel-based extreme learning machines with advanced optimization algorithms. *Environ. Sci. Pollut. Res.* **2022**, *29*, 20496–20516. [CrossRef]
40. Zhao, G.Y.; Suzuki, S.; Deng, J.H.; Fujita, M. Machine learning estimation of biodegradable organic matter concentrations in municipal wastewater. *J. Environ. Manag.* **2022**, *323*, 116191. [CrossRef]
41. Kim, Y.; Park, S.; Oh, S. Machine Learning Approach Reveals the Assembly of Activated Sludge Microbiome with Different Carbon Sources during Microcosm Startup. *Microorganisms* **2021**, *9*, 1387. [CrossRef]
42. American Public Health Association. *Standard Methods for the Examination of Water and Wastewater*; American Public Health Association: Washington, DC, USA, 2005; ISBN 978-0-8755-3287-5.
43. Murphy, K.R.; Stedmon, C.A.; Waite, T.D.; Ruiz, G.M. Distinguishing between terrestrial and autochthonous organic matter sources in marine environments using fluorescence spectroscopy. *Mar. Chem.* **2008**, *108*, 40–58. [CrossRef]
44. Bahramian, M.; Dereli, R.K.; Zhao, W.; Giberti, M.; Casey, E. Data to intelligence: The role of data-driven models in wastewater treatment. *Expert Syst. Appl.* **2023**, *217*, 119453. [CrossRef]
45. Temel, F.A.; Yolcu, O.C.; Turan, N.G. Artificial intelligence and machine learning approaches in composting process: A review. *Bioresour. Technol.* **2023**, *370*, 128539. [CrossRef]
46. Adibimanesh, B.; Polesek-Karczewska, S.; Bagherzadeh, F.; Szczuko, P.; Shafighfard, T. Energy consumption optimization in wastewater treatment plants: Machine learning for monitoring incineration of sewage sludge. *Sustain. Energy Technol. Assess.* **2023**, *56*, 103040. [CrossRef]
47. Dimple; Singh, P.K.; Rajput, J.; Kumar, D.; Gaddikeri, V.; Elbeltagi, A. Combination of discretization regression with data-driven algorithms for modeling irrigation water quality indices. *Ecol. Inform.* **2023**, *75*, 102093. [CrossRef]

48. Zhang, W.; Chen, Q.; Chen, J.; Xu, D.; Zhan, H.; Peng, H.; Pan, J.; Vlaskin, M.; Leng, L.; Li, H. Machine learning for hydrothermal treatment of biomass: A review. *Bioresour. Technol.* **2023**, *370*, 128547. [CrossRef]
49. Suwa, T.; Fujiu, M.; Morisaki, Y.; Fukuoka, T. Analysis of Estimation of Soundness and Deterioration Factors of Sewage Pipes Using Machine Learning. *Sustainability* **2023**, *15*, 16081. [CrossRef]
50. Abba, S.I.; Yassin, M.A.; Mubarak, A.S.; Shah, S.M.H.; Usman, J.; Oudah, A.Y.; Naganna, S.R.; Aljundi, I.H. Drinking Water Resources Suitability Assessment Based on Pollution Index of Groundwater Using Improved Explainable Artificial Intelligence. *Sustainability* **2023**, *15*, 15655. [CrossRef]
51. Wang, S.; Zhang, X.; Wang, Z.W.; Li, X.; Ma, J. In-depth characterization of secondary effluent from a municipal wastewater treatment plant located in Northern China for advanced treatment. *Water Sci. Technol.* **2014**, *69*, 1482–1488. [CrossRef] [PubMed]
52. Park, M.H.; Lee, T.H.; Lee, B.M.; Hur, J.; Park, D.H. Spectroscopic and Chromatographic Characterization of Wastewater Organic Matter from a Biological Treatment Plant. *Sensors* **2010**, *10*, 254–265. [CrossRef]
53. Xue, S.; Jin, W.; Zhang, Z.; Liu, H. Reductions of dissolved organic matter and disinfection by-product precursors in full-scale wastewater treatment plants in winter. *Chemosphere* **2017**, *179*, 395–404. [CrossRef]
54. Zhang, Q.H.; Jin, P.K.; Ngo, H.H.; Shi, X.; Guo, W.S.; Yang, S.J.; Wang, X.C.; Wang, X.; Dzakpasu, M.; Yang, W.N.; et al. Transformation and utilization of slowly biodegradable organic matters in biological sewage treatment of anaerobic anoxic oxic systems. *Bioresour. Technol.* **2016**, *218*, 53–61. [CrossRef]
55. Zhao, C.; Wang, Z.; Wang, C.; Li, X.; Wang, C.C. Photocatalytic degradation of DOM in urban stormwater runoff with TiO_2 nanoparticles under UV light irradiation: EEM-PARAFAC analysis and influence of co-existing inorganic ions. *Environ. Pollut.* **2018**, *243*, 177–188. [CrossRef]
56. Zeng, Z.; Zheng, P.; Ding, A.; Zhang, M.; Abbas, G.; Li, W. Source analysis of organic matter in swine wastewater after anaerobic digestion with EEM-PARAFAC. *Environ. Sci. Pollut. Res.* **2017**, *24*, 6770–6778. [CrossRef]
57. Murphy, K.R.; Hambly, A.; Singh, S.; Henderson, R.K.; Baker, A.; Stuetz, R.; Khan, S.J. Organic Matter Fluorescence in Municipal Water Recycling Schemes: Toward a Unified PARAFAC Model. *Environ. Sci. Technol.* **2011**, *45*, 2909–2916. [CrossRef]
58. Yu, H.; Qu, F.; Sun, L.; Liang, H.; Han, Z.; Chang, H.; Shao, S.; Li, G. Relationship between soluble microbial products (SMP) and effluent organic matter (EfOM): Characterized by fluorescence excitation emission matrix coupled with parallel factor analysis. *Chemosphere* **2015**, *121*, 101–109. [CrossRef]
59. Ren, W.; Wu, X.; Ge, X.; Lin, G.; Zhou, M.; Long, Z.; Yu, X.; Tian, W. Characteristics of dissolved organic matter in lakes with different eutrophic levels in southeastern Hubei Province, China. *J. Oceanol. Limnol.* **2021**, *39*, 1256–1276. [CrossRef]
60. Chen, B.; Huang, W.; Ma, S.; Feng, M.; Liu, C.; Gu, X.; Chen, K. Characterization of Chromophoric Dissolved Organic Matter in the Littoral Zones of Eutrophic Lakes Taihu and Hongze during the Algal Bloom Season. *Water* **2018**, *10*, 861. [CrossRef]
61. Derrien, M.; Lee, M.H.; Choi, K.; Lee, K.S.; Hur, J. Tracking the evolution of particulate organic matter sources during summer storm events via end-member mixing analysis based on spectroscopic proxies. *Chemosphere* **2020**, *252*, 126445. [CrossRef] [PubMed]
62. Rodriguez-Vidal, F.J.; Garcia-Valverde, M.; Ortega-Azabache, B.; Gonzalez-Martinez, A.; Bellido-Fernandez, A. Characterization of urban and industrial wastewaters using excitation-emission matrix (EEM) fluorescence: Searching for specific fingerprints. *J. Environ. Manag.* **2020**, *263*, 110396. [CrossRef]
63. Wang, Y.; Hu, Y.; Yang, C.; Wang, Q.; Jiang, D. Variations of DOM quantity and compositions along WWTPs-river-lake continuum: Implications for watershed for environmental management. *Chemosphere* **2019**, *218*, 468–476. [CrossRef]
64. Wei, L.L.; Zhao, Q.L.; Xue, S.; Jia, T.; Tang, F.; You, P.Y. Behavior and characteristics of DOM during a laboratory-scale horizontal subsurface flow wetland treatment: Effect of DOM derived from leaves and roots. *Ecol. Eng.* **2009**, *35*, 1405–1414. [CrossRef]
65. Jathan, Y.; Pagilla, K.R.; Marchand, E.A. Understanding the influence of dissolved organic nitrogen characteristics on enhanced coagulation performance for water reuse. *Chemosphere* **2023**, *337*, 139384. [CrossRef]
66. Eom, H.; Borgatti, D.; Paerl, H.W.; Park, C. Formation of Low-Molecular-Weight Dissolved Organic Nitrogen in Predenitrification Biological Nutrient Removal Systems and Its Impact on Eutrophication in Coastal Waters. *Environ. Sci. Technol.* **2017**, *51*, 3776–3783. [CrossRef]
67. Maqbool, T.; Cho, J.; Hur, J. Spectroscopic descriptors for dynamic changes of soluble microbial products from activated sludge at different biomass growth phases under prolonged starvation. *Water Res.* **2017**, *123*, 751–760. [CrossRef]
68. Komatsu, K.; Onodera, T.; Kohzu, A.; Syutsubo, K.; Imai, A. Characterization of dissolved organic matter in wastewater during aerobic, anaerobic, and anoxic treatment processes by molecular size and fluorescence analyses. *Water Res.* **2020**, *171*, 115459. [CrossRef] [PubMed]
69. Ding, X.; Wei, D.; Guo, W.; Wang, B.; Meng, Z.; Feng, R.; Du, B.; Wei, Q. Biological denitrification in an anoxic sequencing batch biofilm reactor: Performance evaluation, nitrous oxide emission and microbial community. *Bioresour. Technol.* **2019**, *285*, 121359. [CrossRef] [PubMed]
70. Gao, L.; Han, F.; Zhang, X.; Liu, B.; Fan, D.; Sun, X.; Zhang, Y.; Yan, L.; Wei, D. Simultaneous nitrate and dissolved organic matter removal from wastewater treatment plant effluent in a solid-phase denitrification biofilm reactor. *Bioresour. Technol.* **2020**, *314*, 123714. [CrossRef] [PubMed]
71. Li, J.; Wei, J.; Ngo, H.H.; Guo, W.; Liu, H.; Du, B.; Wei, Q.; Wei, D. Characterization of soluble microbial products in a partial nitrification sequencing batch biofilm reactor treating high ammonia nitrogen wastewater. *Bioresour. Technol.* **2018**, *249*, 241–246. [CrossRef] [PubMed]

72. Liu, Y.; Wei, D.; Xu, W.; Feng, R.; Du, B.; Wei, Q. Nitrogen removal in a combined aerobic granular sludge and solid-phase biological denitrification system: System evaluation and community structure. *Bioresour. Technol.* **2019**, *288*, 121504. [CrossRef] [PubMed]
73. Yang, Y.; Yu, Y.; Zhou, J.; Xi, H.; Wang, C.; Zhou, Y.; Fu, L.; Qi, Y.; Yuan, Y. Characterization and Analysis of Acetaldehyde Wastewater by Molecular Weight Distribution, Hydrophilicity, and Chemical Composition. *Sustainability* **2022**, *14*, 6540. [CrossRef]
74. Zhao, C.; Gong, J.; Zeng, Q.; Yang, M.; Wang, Y. Landscape Pattern Evolution Processes and the Driving Forces in the Wetlands of Lake Baiyangdian. *Sustainability* **2021**, *13*, 9747. [CrossRef]
75. Matomela, N.; Li, T.; Zhang, P.; Ikhumhen, H.O.; Lopes, N.D.R. Role of Landscape and Land-Use Transformation on Nonpoint Source Pollution and Runoff Distribution in the Dongsheng Basin, China. *Sustainability* **2023**, *15*, 8325. [CrossRef]
76. Ma, J.; Ding, Y.; Cheng, J.C.P.; Jiang, F.; Xu, Z. Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques. *Water Res.* **2020**, *170*, 115350. [CrossRef] [PubMed]
77. Nguyen, D.-V.; Park, J.; Lee, H.; Han, T.; Wu, D. Assessing industrial wastewater effluent toxicity using boosting algorithms in machine learning: A case study on ecotoxicity prediction and control strategy development. *Environ. Pollut.* **2024**, *341*, 123017. [CrossRef] [PubMed]
78. Zhang, Y.; Sun, Q.; Liu, J.; Petrosian, O. Long-Term Forecasting of Air Pollution Particulate Matter ($PM_{2.5}$) and Analysis of Influencing Factors. *Sustainability* **2024**, *16*, 19. [CrossRef]
79. Liu, X.; Wang, D.; Chen, Z.; Wei, W.; Mannina, G.; Ni, B.J. Advances in pretreatment strategies to enhance the biodegradability of waste activated sludge for the conversion of refractory substances. *Bioresour. Technol.* **2022**, *362*, 127804. [CrossRef]
80. Li, H.; Zhou, B.; Xu, X.; Huo, R.; Zhou, T.; Dong, X.; Ye, C.; Li, T.; Xie, L.; Pang, W. The insightful water quality analysis and predictive model establishment via machine learning in dual-source drinking water distribution system. *Environ. Res.* **2024**, *250*, 118474. [CrossRef]
81. Zhou, B.; Li, H.; Wang, Z.; Huang, H.; Wang, Y.; Yang, R.; Huo, R.; Xu, X.; Zhou, T.; Dong, X. Prediction of phosphate adsorption amount, capacity and kinetics via machine learning: A generally physical-based process and proposed strategy of using descriptive text messages to enrich datasets. *Chem. Eng. J.* **2024**, *479*, 147503. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.