Liu *et al. Microbiome* (2023) 11:93
https://doi.org/10.1186/s40168-023-01519-9

Microbiome

RESEARCH

Open Access

Predicting microbial community compositions in wastewater treatment plants using artificial neural networks

Xiaonan Liu[1], Yong Nie[1*] and Xiao-Lei Wu[1,2,3*]

Abstract

Background Activated sludge (AS) of wastewater treatment plants (WWTPs) is one of the world's largest artificial microbial ecosystems and the microbial community of the AS system is closely related to WWTPs' performance. However, how to predict its community structure is still unclear.

Results Here, we used artificial neural networks (ANN) to predict the microbial compositions of AS systems collected from WWTPs located worldwide. The predictive accuracy $R^2_{1:1}$ of the Shannon–Wiener index reached 60.42%, and the average $R^2_{1:1}$ of amplicon sequence variants (ASVs) appearing in at least 10% of samples and core taxa were 35.09% and 42.99%, respectively. We also found that the predictability of ASVs was significantly positively correlated with their relative abundance and occurrence frequency, but significantly negatively correlated with potential migration rate. The typical functional groups such as nitrifiers, denitrifiers, polyphosphate-accumulating organisms (PAOs), glycogen-accumulating organisms (GAOs), and filamentous organisms in AS systems could also be well recovered using ANN models, with $R^2_{1:1}$ ranging from 32.62% to 56.81%. Furthermore, we found that whether industry wastewater source contained in inflow (IndConInf) had good predictive abilities, although its correlation with ASVs in the Mantel test analysis was weak, which suggested important factors that cannot be identified using traditional methods may be highlighted by the ANN model.

Conclusions We demonstrated that the microbial compositions and major functional groups of AS systems are predictable using our approach, and IndConInf has a significant impact on the prediction. Our results provide a better understanding of the factors affecting AS communities through the prediction of the microbial community of AS systems, which could lead to insights for improved operating parameters and control of community structure.

Keywords Activated sludge, Artificial neural networks, Prediction, Microbial compositions, Functional groups

Background

With the increasing expansion of urbanization, about 360 billion m^3 of wastewater is produced every year globally [1]. The activated sludge (AS) system in wastewater treatment plants (WWTPs) is at the heart of current sewage treatment technology [2]. Microorganisms treat almost 60% of this wastewater in AS systems before release [3]. This process relies on the degradation of organic compounds, biotransformation of toxic substances, and removal of pathogens by diverse microorganisms [4–6]. Thus, the microbial communities present in these

*Correspondence:
Yong Nie
nieyong@pku.edu.cn
Xiao-Lei Wu
xiaolei_wu@pku.edu.cn
[1] College of Engineering, Peking University, Beijing 100871, China
[2] Institute of Ocean Research, Peking University, Beijing 100871, China
[3] Institute of Ecology, Peking University, Beijing 100871, China

systems determine their performance [7]. Predicting the microbial communities of AS systems and exploring factors that influence them will provide reasonable suggestions for the design, optimization, and stable operation of sewage treatment systems [8–11]. However, because wastewater always contains a multiplicity of resources, the AS system exhibits an enormous microbial diversity and varies greatly worldwide. The global activated sludge community encompasses about 1 billion bacterial phylotypes, with a small global core bacterial community consisting of only 28 operational taxonomic units (OTUs) [3]. The overwhelming taxonomic diversity and variability of microbial communities in AS systems pose a significant challenge for accurate modeling and predicting their structure and function. It is still unclear how we can predict the microbial communities in the AS systems of WWTPs according to the design parameters and environmental data.

The AS system contains high biomass and microbial diversity [3, 12], and predicting the microbial community is complicated by diverse factors. For example, AS systems treating municipal and industrial wastewater harbor distinct microbial communities [13, 14], suggesting that the type of wastewater impacts microbial composition. The influent biodegradability [biological oxygen demand/chemical oxygen demand (B/C ratio)] also plays an essential role in shaping the AS microbial community. A low or high B/C ratio may lead to low microbial diversity and pollutant removal loading [15], indicating that the impact of the B/C ratio on community structure may be nonlinear. Recently, the integration of high-throughput sequencing and multivariate statistical analysis indicated that the microbial communities of AS systems are significantly correlated with multiple factors, such as location, geographical distance, dissolved oxygen (DO), temperature, hydraulic retention time (HRT), sludge retention time (SRT), inflow and effluent of chemical oxygen demand (COD), total nitrogen (TN), total phosphorus (TP) [16, 17]. The combined influence of multiple environmental factors has made it difficult to predict the microbial compositions in AS systems and thus has become an obstacle to guiding the operation of WWTPs.

Mechanism-based kinetic models, such as the Monod equation, Lotka-Volterra model, and individual-based dynamic model can predict the structure of microbial communities based on specific growth and interaction mechanisms under given conditions [18–21]. However, these models are limited in their capacity to generalize to complex natural communities due to simplified growth or interaction assumptions. Multiple linear regression models can predict microbial community structure from multiple environmental factors. A previous study predicted bacterial and fungal groups in a soil microbial community from typical soil environmental factors [C and N concentrations, pH, mean annual temperature (MAT), mean annual precipitation (MAP) and net primary productivity (NPP), etc.] using this method [22]. However, since multiple regression models ignore the interaction effects of environmental factors and nonlinear relationships, the predictability of the microbial taxa in that study was at most no more than 60%. The AS system is affected by multiple cross-complex factors, including geographical factors, design and operation parameters, and physicochemical parameters, and multiple regression analysis is not enough to capture this complex relationship. In addition, no attention has been paid to the regularity of predictability of microbial taxa in previous studies, which is essential for a deeper understanding and control of microbial community structure.

Artificial neural network (ANN) is a machine learning method for the automatic and quantitative learning of a suitable relationship without any specific assumptions and guiding system optimization [23]. The ANN is an ideal alternative to model these complex relationships between microbial communities and environmental variables as this method is better suited to account for the non-linear associations between variables and the interactions among predictors [24]. ANNs have helped researchers to successfully analyze the relationship between environmental factors and microbial community structure in many ecosystems [24–26], while the relevant applications of activated sludge systems are still lacking. Considering the strong ability of the ANN model to predict complex systems, we hypothesized that the ANN model can predict the microbial community structure of AS system.

Here, we used ANN models and environment data to predict the microbial community structure of AS systems from global wastewater treatment plants. We analyzed the predictability of different taxa and the effects of environmental factors on the prediction. These analyses deepened our understanding of the microbial community of AS systems, provided reasonable suggestions for accurately predicting major functional groups, and provided a theoretical basis for better design and operating parameters, and to control community structure.

Results

Overview of microbial community structure in AS systems

By preprocessing 777 (no data leakage) activated sludge samples from 269 wastewater treatment plants located in 23 countries across 6 continents using the QIIME2 pipeline, we obtained the basic information of microbial community structure in AS systems. Specifically, the Shannon–Wiener index ranged from 2.90 to 6.41 (Fig. 1a), Pielou's evenness index ranged from 0.50 to

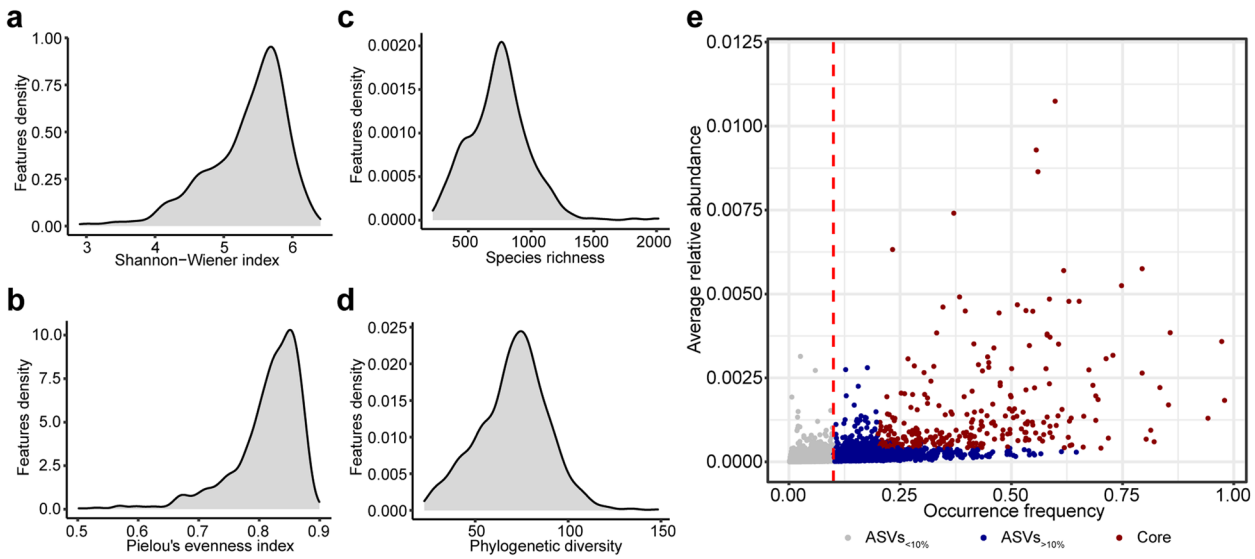

Fig. 1 Overview of microbial community structure in AS system. Distribution of Shannon–Wiener index (**a**), Pielou's evenness index (**b**), species richness (**c**), and Faith's phylogenetic diversity (**d**). **e**. Occurrence frequency and average relative abundance distribution of all ASVs in the AS system

0.90 (Fig. 1b), species richness ranged from 217 to 2014 (Fig. 1c), and Faith's phylogenetic diversity ranged from 22.58 to 148.33 (Fig. 1d). Detailed information about alpha diversity is provided in Table S1.

In addition, we analyzed the distribution features of the average relative abundance and occurrence frequency of the ASVs. The results showed that ASVs in the AS systems were dominated by low relative abundance (Fig. 1e), which is in line with the general ecological environment [27]. In this study, we only predicted 1493 ASVs appeared in at least 10% of samples, of which 290 belonged to the core ASVs (Fig. 1e), which was defined as overall abundant, ubiquitous, and frequently abundant ASVs.

Alpha-diversities of AS systems can be predicted by ANN models

Predictability of alpha-diversities

To obtain an overall prediction of AS community structure, we first constructed predictive models for different alpha diversity indices, including the Shannon–Wiener index, Pielou's evenness index, species richness, and Faith's phylogenetic diversity. Here, the predictive accuracy is measured relative to the 1:1 observed-predicted line (rather than a best-fit line), named $R^2_{1:1}$, so accuracy assessments are both qualitative and quantitative [22]. By comparing the observed and predicted alpha diversities in test sets, we found that predictive accuracies $R^2_{1:1}$ of the Shannon–Wiener index (Fig. 2b), Pielou's evenness index (Fig. 2c), species richness (Fig. 2d), and Faith's phylogenetic diversity (Fig. 2e) were 60.42%, 54.11%, 49.92%, and 60.37%, respectively.

Comparing the predictability of different alpha diversity indices, we found that the Shannon–Wiener index and Pielou's evenness index were more predictable than species richness, which may be related to the environmental sensitivity of species evenness. Species evenness has previously been reported to be more sensitive to human activity and environmental changes than richness because environmental conditions may significantly affect ecosystems long before a species is threatened by extinction [28]. In addition, the predictive accuracy of phylogenetic diversity was also higher than species richness, reflecting that species' evolutionary history may be influenced by environmental factors surrounding the microbial community.

Environmental factors important for predicting alpha-diversities

During the model training process for predicting alpha-diversities of AS systems, an importance weight value was assigned to each environmental factor by Garson's connection weight method [29]. The factors with higher importance weights were more informative when the model was used to predict alpha diversities.

To assess the importance of different environmental factors in predicting alpha-diversities of AS microbial communities, we ranked the average importance weights of environmental factors in different predictive models in descending order (Additional file 2: Figure S1). The results showed that DO was most important for predicting the Shannon–Wiener and Pielou's evenness indices, but inflow-relatedIndConInf was most important

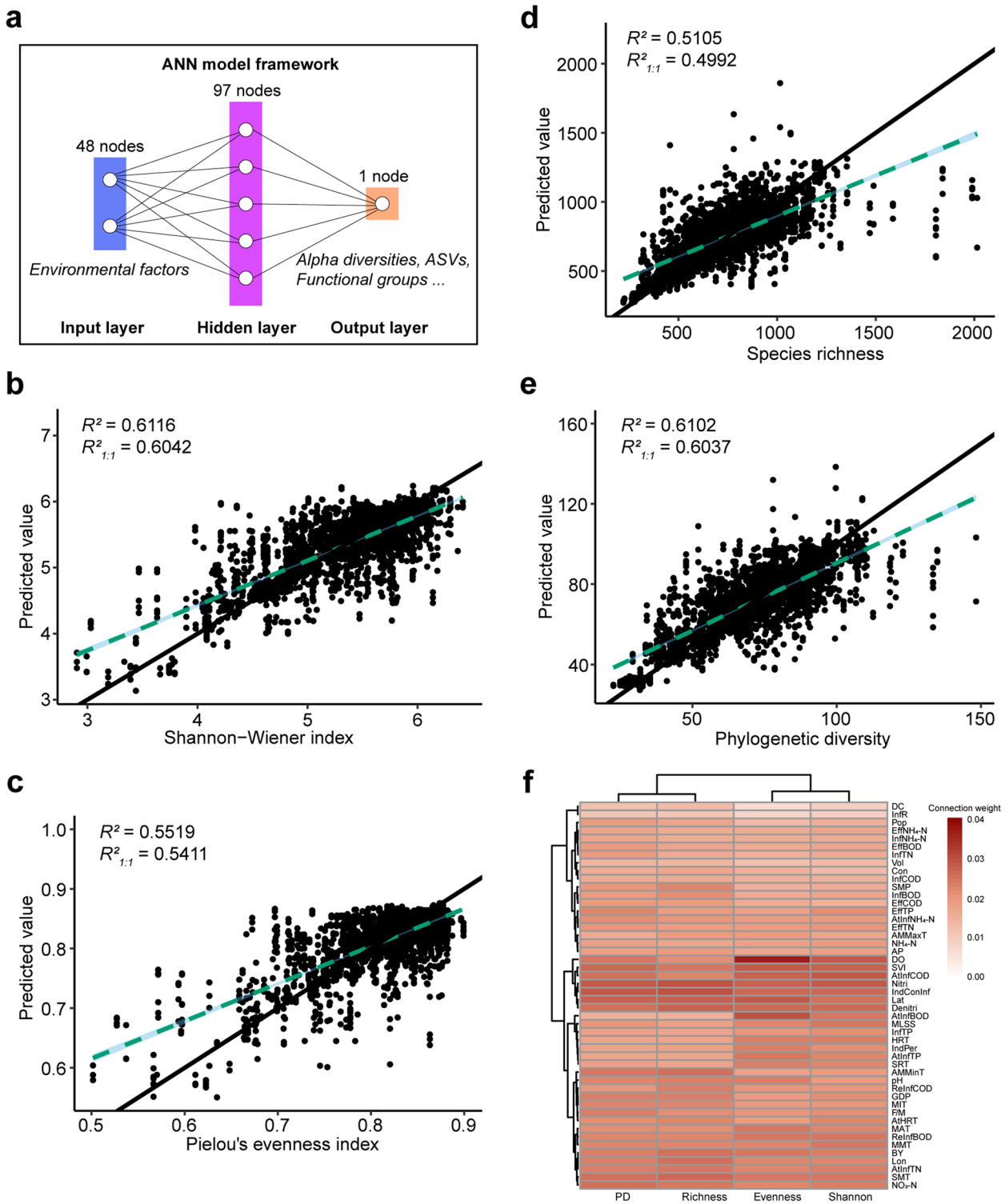

Fig. 2 Prediction of alpha diversity. **a.** The framework of ANN models: input data (blue), output data (red), and a predictive model trained to compute output data from input data (purple). Correlations between observed and predicted values of Shannon–Wiener index (**b**), Pielou'e evenness index (**c**), species richness (**d**), and Faith's phylogenetic diversity (**e**). The 1:1 relationship is shown as a solid black line, and the best fit is shown as the dashed light blue line. The blue-shaded region represents the 95% confidence interval for the best-fit line. We reported the R^2 value of the best-fit line between predicted and observed and the R^2 observations relative to the 1:1 line. **f.** Heatmap of importance weights of environmental factors in alpha diversity predictive models

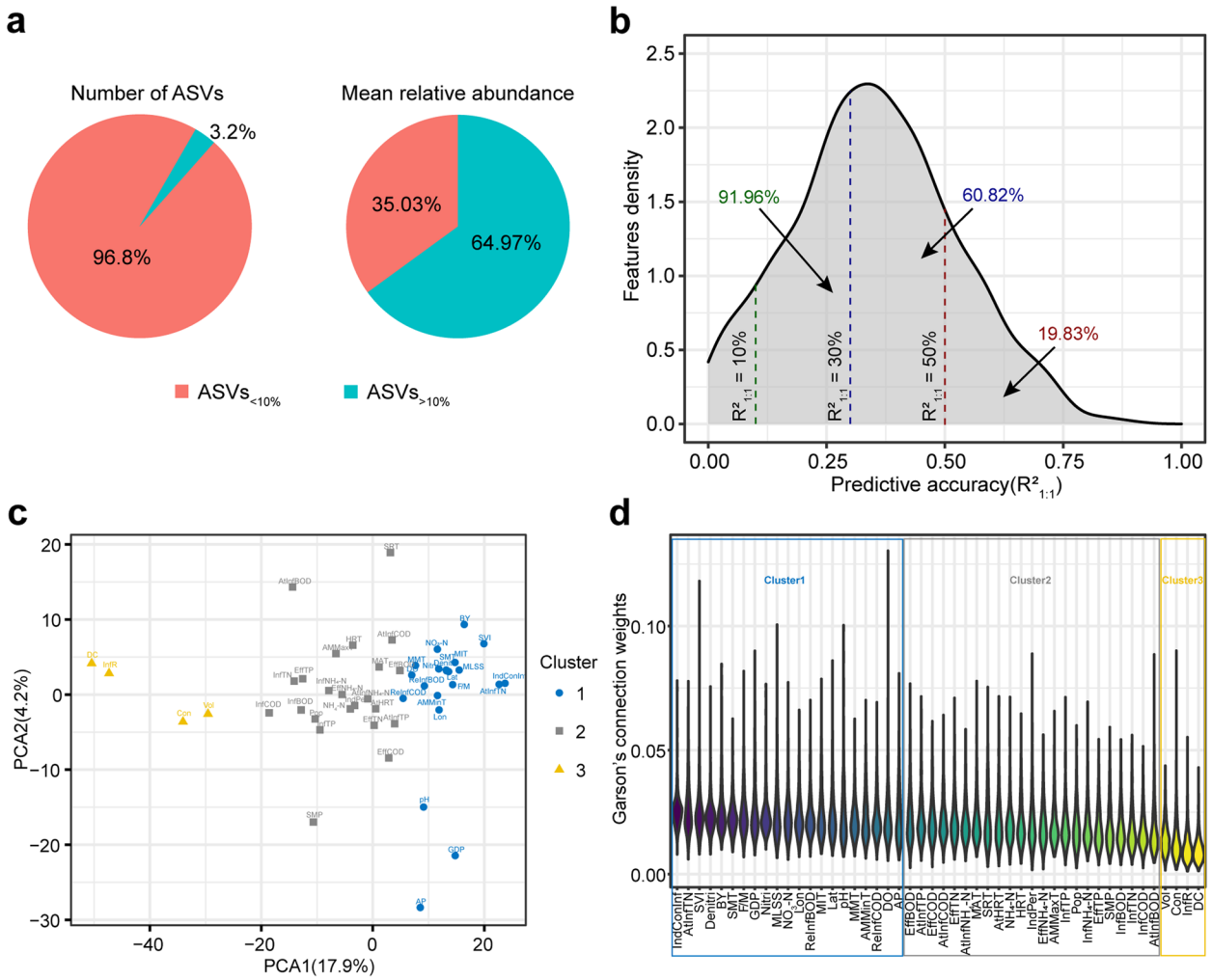

Fig. 3 Prediction of the relative abundance of ASVs$_{>10\%}$. **a.** Percentage of ASV number and relative abundance of ASVs$_{<10\%}$ versus ASVs$_{>10\%}$. **b.** Distribution of the predictive accuracy of ASVs$_{>10\%}$. The dark green, dark blue, and dark red text represents the proportion of ASVs with prediction accuracy exceeding 10%, 30%, and 50%, respectively. **c.** Principal component analysis (PCA) of environmental factors colored by k-means clusters. **d.** Ranking of environmental factors in descending order of median importance weights

for predicting species richness and Faith's phylogenetic diversity. Climatic condition latitude (Lat), design-related N removal process [nitrification (Nitri) and denitrification (Denitri)], COD in the inflow of aeration tank (AtInfCOD), and the sludge volume index (SVI) were also environmental factors with high average importance weights for predicting alpha diversities (Fig. 2f).

Assessment of the predictivity of community structure using the ANN model

Predictability of the relative abundances of ASVs

To obtain a deep prediction of AS community structure, we predicted the relative abundance of ASVs in AS systems. We constructed predictive models for the 1493 ASVs found in more than 10% of samples (ASVs$_{>10\%}$), which accounted for 3.2% of the total ASVs and 64.97 ± 0.54% (mean ± SEM) of the relative abundance in AS samples (Fig. 3a). The results showed that the average predictive accuracy $R^2_{1:1}$ of ASVs$_{>10\%}$ was 35.09% (Table S2). Further, we found that 19.83% of ASVs$_{>10\%}$ could be predicted with $R^2_{1:1}$ over 50%, 60.82% of ASVs$_{>10\%}$ could be predicted with $R^2_{1:1}$ over 30%, and 91.96% of ASVs$_{>10\%}$ could be predicted with $R^2_{1:1}$ over 10% (Fig. 3b).

In addition, we also predicted the structures of the microbial communities of the test samples, by recovering the ASVs$_{>10\%}$ subcommunity of each sample in its entirety [25]. Here, we refer to the observed values of ASVs$_{>10\%}$ subcommunities in different test samples as "observed communities", and the corresponding predicted values as "predicted communities". By comparing

the intra-group and inter-group differences between the predicted and observed communities, we found that the Bray–Curtis similarity between intra-groups was significantly higher than that between inter-groups (Additional file 2: Figure S2a). This result also proved that the ANN model could predict the microbial community structure of the AS system from an overall perspective.

Furthermore, we predicted microbial taxa at different taxonomic levels and found that microbial community structure had high average predictive accuracy ranging from 33.32% to 41.6% at all taxonomic levels (Additional file 2: Figure S2b). The predictive accuracy $R^2_{1:1}$ of the three most abundant phyla (*Proteobacteria*, *Bacteroidota,* and *Myxococcota*) in the AS system were 64.54%, 55.37%, and 59.04%, respectively. The three most abundant orders *Burkholderiales*, *Chitinophagales,* and *Pseudomonadales* could be predicted with $R^2_{1:1}$ of 63.89%, 56.19%, and 42.81%, respectively (Table S3).

Importance of environmental factors in the prediction of ASVs

During the model training process for predicting abundances of ASVs, an importance weight value was also assigned to each environmental factor as above (Table S4). By displaying the importance weights of environmental factors in different ASVs predictive models, we found that environmental factors had different weights in predicting different ASVs (Additional file 2: Figure S3a). Further, we clustered the environmental factors into three clusters according to their importance weights using the k-means clustering algorithm and displayed them using principal components analysis (PCA) (Fig. 3c). We found that these three clusters corresponded to three parts divided by the median of importance weights in descending order (Fig. 3d). This result showed that environmental factors of cluster1, which included climatic condition sampling moment temperature (SMT), design and operation parameters year of plant build (BY) and Denitri, inflow conditions IndConInf and total nitrogen in the inflow of aeration tank (AtInfTN), and physicochemical properties SVI, etc., contributed the most to the prediction of community structure, cluster2 was second, and cluster3 was the least important group of factors in predicting community structure (Table S5).

We then wondered what influences the importance weights of environmental factors in predicting microbial taxa. Before constructing the predictive model, we performed a Mantel test analysis on the correlation between the $ASVs_{>10\%}$ subcommunity and ecological environment factors (Table S5). The correlation analysis showed that environmental factors significantly associated with the $ASVs_{>10\%}$ subcommunity included climate conditions Lat, longitude (Lon), MAT, the annual mean of daily maximum temperature (AMMinT), sampling month precipitation (SMP), and GDP, design and operation parameter Nitri, and physicochemical property mixed liquid temperature (MIT) (Pearson's $\rho > 0.2$, $p < 0.01$). By comparing the importance of environmental factors in predicting community structure and the correlation between environmental factors and community structure, we found that some of the environmental factors that had high importance weights in many predictive models were not strongly correlated with the $ASVs_{>10\%}$ subcommunity (Additional file 2: Figure S3). For example, inflow conditions IndConInf and AtInfTN, which were important for predicting the relative abundance of ASVs, did not significantly correlate with the $ASVs_{>10}$ sub-community. However, despite these differences, there was a significant positive correlation between the importance weights of environmental factors and their correlation coefficients with the $ASVs_{>10\%}$ subcommunity (Additional file 2: Figure S4a, $R^2 = 0.1271$, $p < 0.05$). These results showed that in addition to the correlation analysis, importance weights analysis in ANN predictive models also helped to expand the range of environmental factors that should be paid attention to when exploring the performance of WWTPs.

In addition, we also analyzed the influence of the distribution of environmental factors on their weights. The result showed that both the skewness (Additional file 2: Figure S4b; $R^2 = 0.6268$, $p < 0.001$) and kurtosis (Additional file 2: Figure S4c; $R^2 = 0.7106$, $p < 0.001$) of normalized environmental factors were significantly negatively correlated with their average importance weights in predicting ASVs. This suggested that environmental factors of low skewness and low kurtosis may be more important in predicting community structure.

Characteristics of ASVs with high predictabilities

ASVs with higher relative abundances and occurrence frequencies can be better predicted using the ANN model

To investigate the correlation between the predictability of ASVs and their distribution features, we compared the predictability of ASVs with different relative abundances and frequencies. The correlation analysis between the predictive accuracy $R^2_{1:1}$ of all 1493 $ASVs_{>10\%}$ and their relative abundances showed that the $R^2_{1:1}$ of an ASV was significantly positively correlated with its relative abundance (Fig. 4a; $R^2 = 0.05279$, $P < 0.001$), indicating that the high predictability of an ASV may be related to its high relative abundance. Furthermore, we found that the $R^2_{1:1}$ of an ASV was slightly positively associated with its occurrence frequency at significant levels (Fig. 4b; $R^2 = 0.02602$, $P < 0.001$), indicating that the high predictability of abundant $ASV_{>10\%}$ may also be related to its high occurrence frequency. Further, we grouped

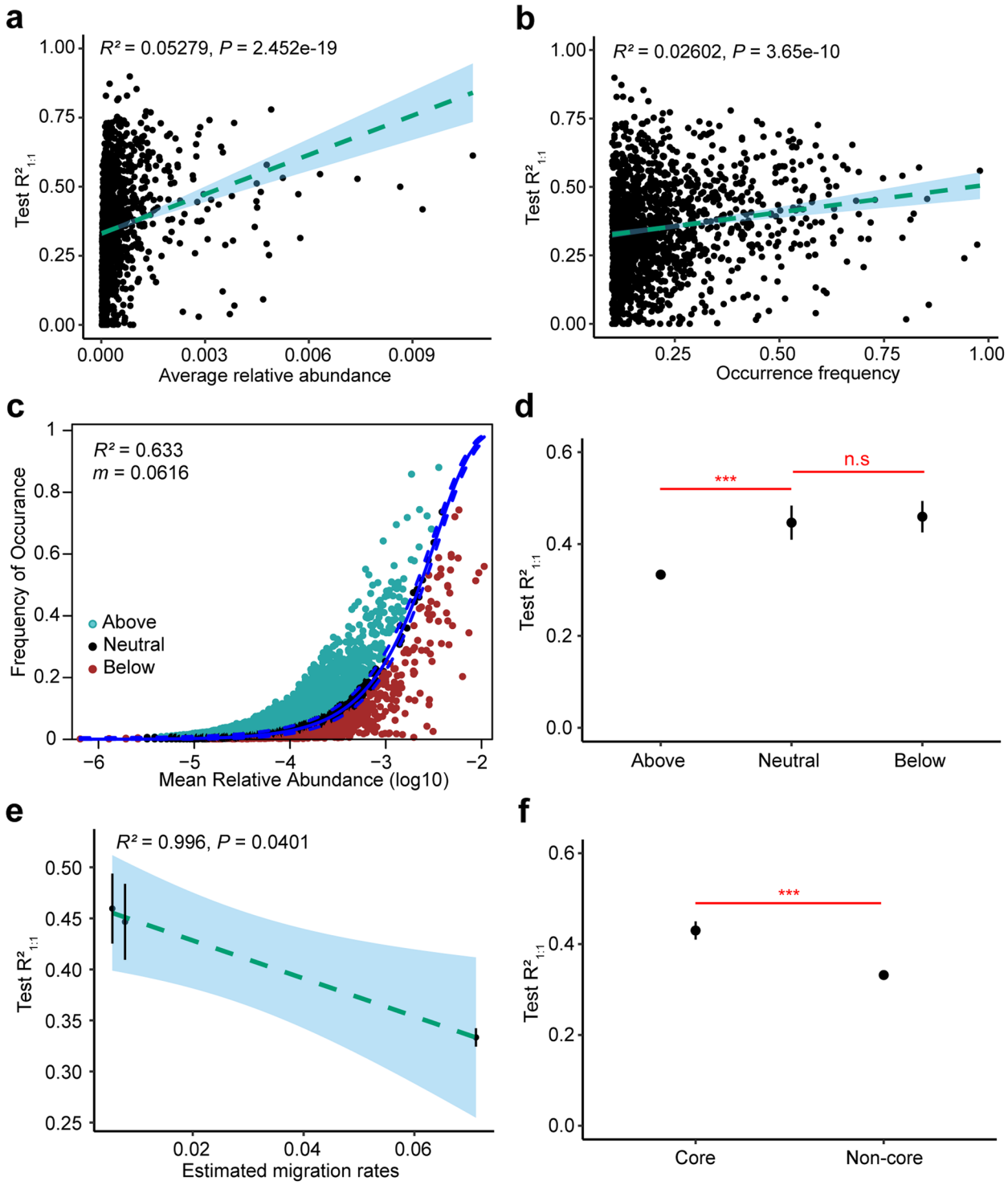

Fig. 4 Distribution features and predictability of $ASV_{>10\%}$. **a.** Correlation of test $R^2_{1:1}$ with the average relative abundance. **b.** Correlation of test $R^2_{1:1}$ with the occurrence frequency. **c.** Fit of the neutral community model (NCM) of AS system community assembly. The solid blue lines indicate the best fit to the NCM as in Sloan et al. [31], and the dashed blue lines represent 95% confidence intervals around the model prediction. R^2 indicates the fit to this model, and m indicates the estimated migration rate. **d.** The test $R^2_{1:1}$ of above, neutral, and below partitions. **e.** Correlation of test $R^2_{1:1}$ with the estimated migration rate. The data was provided by the results of different partitions. **f.** The test $R^2_{1:1}$ of core and non-core taxa. Statistical analysis was performed using a two-sample Student's *t*-test: ***, $p < 0.001$; n.s, $p > 0.05$, no significance

ASVs based on their relative abundances and occurrence frequencies as in previous studies (see more details in Additional file 1), and the results showed that $ASVs_{>10\%}$ with a high relative abundance and high occurrence frequency were significantly more predictable than those with low relative abundance (Additional file 2: Figure S5a) and low occurrence frequency (Additional file 2: Figure S5b), which is consistent with the previous result. It is worth mentioning that the occurrence frequency of an ASV has a significant positive correlation with its relative abundance (Additional file 2: Figure S5c, $R^2 = 0.2978$, $P < 0.001$), supporting that high relative abundance and high occurrence frequency can corroborate each other in their contribution to predictability.

Previous studies had demonstrated that rare taxa were more dynamic than abundant taxa [30], so we wondered whether a taxon's predictability was related to its variability across samples. To explore this question, we analyzed the relationship between $R^2_{1:1}$ of the ASVs and their coefficients of variation and found that the predictive accuracy of an ASV was significantly negatively correlated with its coefficient of variation (Additional file 2: Figure S5d; $R^2 = 0.01946$, $P < 0.001$), implying that taxa with higher variability were less predictable.

The predictability of an ASV decreases as its potential migration rate increases

Previous studies have demonstrated that the process of community assembly is closely related to its predictability [22, 32, 33], so we explored the association between microbial community assembly mechanisms in AS systems and the predictability of the corresponding taxa. By neutral community model (NCM) model fitting, we found that stochastic processes explained 63.3% of the microbial community variation in AS systems (Fig. 4c). The $ASVs_{>10\%}$ were subsequently separated into three partitions (above, below, and neutral) depending on their occurrence frequencies and relative abundance. By comparing the distribution features of the three partitions, we found that the relative abundance (Additional file 2: Figure S6a) and occurrence frequency (Additional file 2: Figure S6b) of $ASVs_{>10\%}$ in the below partition were significantly higher than those of the above partition. Further, we found that the predictive accuracy $R^2_{1:1}$ of the below partition was also significantly higher than that of the above partition (Fig. 4d). This result again showed that ASVs with higher relative abundances and occurrence frequencies can be better predicted using ANN models.

In addition, the estimated migration rates of the different partitions assessed by NCM were also different. Points above the fitting curve represent taxa found more frequently than expected, suggesting that they have a higher migration ability and can disperse to more locations. Points below the fitting curve represent taxa found less frequently than expected, suggesting their lower dispersal ability in WWTPs on a global scale or that they have a stronger response to local environmental conditions. The fitting results also confirmed that the taxa in the above partition had the highest estimated migration rates, and the taxa in the below partition had the lowest estimated migration rates (Additional file 2: Figure S7). Further analysis of the relationship between the migration rate and predictability of different partitions, we found that a taxon's predictability had a high negative correlation with its estimated migration rate (Fig. 4e, $R^2 = 0.996$, $P = 0.0401$), indicating that the predictability of an ASV decreased as its potential migration rate increased.

Core taxa of AS systems can be predicted by ANN models

Due to its highly complex organic environment, activated sludge selects a unique core community that does not overlap with the core communities of other habitats [3]. We evaluated the predictabilities of core taxa that were abundant and ubiquitous using ANN models. As defined in Methods, we obtained 290 core ASVs and 1203 non-core ASVs in the $ASVs_{>10\%}$ subcommunity (Additional file 2: Figure S8a). Our analyses found that the relative abundance (Additional file 2: Figure S8b) and occurrence frequency (Additional file 2: Figure S8c) of core taxa were significantly higher than those of non-core taxa, and the estimated migration rate of core taxa was lower than that of non-core taxa (Additional file 2: Figure S8d).

By assessing the predictability of the core taxa, we found more than 37.59% of the core ASVs could be well predicted with an R^2 of over 50%, more than 78.62% could be well predicted with an R^2 of over 30%, and more than 94.48% could be well predicted with an R^2 of over 10%, and the average prediction accuracy was 42.99% (Table S2), which was significantly higher than the non-core taxa (Fig. 4f). Because the core taxa are reported to be closely related to nitrogen removal, phosphorus removal, and even flocculation enhancement of activated sludge [12, 34, 35], the results implied that the ANN model could be used to assess the performance of WWTPs by predicting the dynamics of the core taxa.

Prediction of major functional groups in AS systems

To more directly understand and control the performance of WWTPs, we predicted and analyzed major functional groups of microbial communities in the AS system using ANN models. Referring to the MiDAS4 database, the functional groups in AS systems include nitrogen removal groups (nitrifiers and denitrifiers), phosphorus removal groups (PAOs), and their competitors (GAOs),

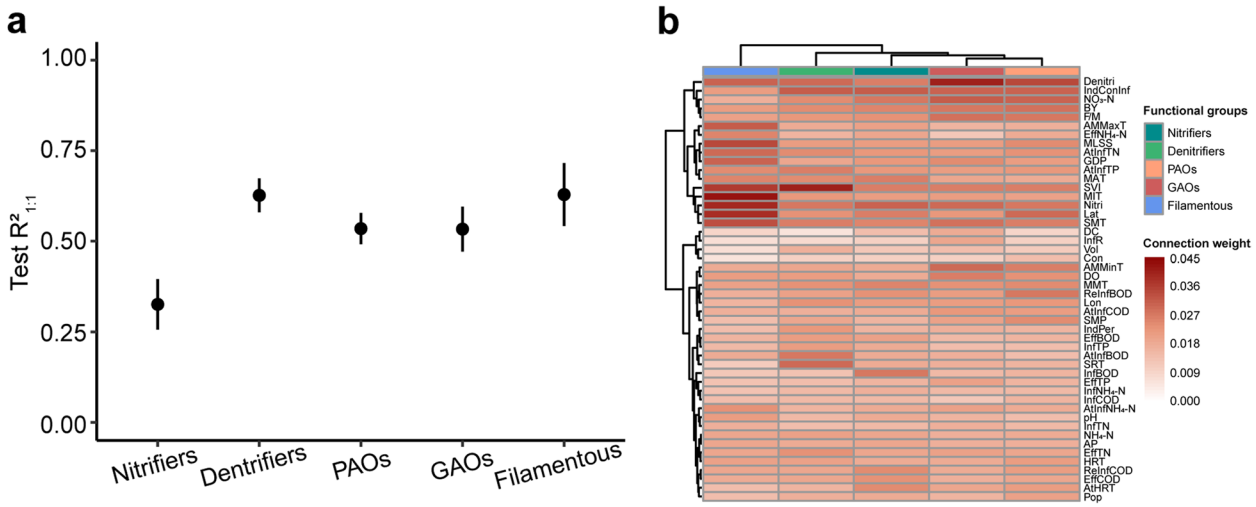

Fig. 5 Prediction of major functional groups. **a.** The test $R^2_{1:1}$ of nitrifiers, denitrifiers, PAOs, GAOs, and filamentous organisms. **b.** Heatmap of importance weights of environmental factors in the predictive models of functional groups. Errors bars in these graphs show the 95% credible intervals of the mean values. Statistical analysis was performed using a two-sample Student's *t*-test: ***, $p < 0.001$; n.s, $p > 0.05$, no significance

and filamentous organisms [36] (Table S6). Then, we calculated the total relative abundance of different functional groups in each sample by summarizing the relative abundances of ASVs from the same functional groups. Finally, we predicted the total relative abundance of each functional group using the ANN model. The results showed that the predictive accuracy $R^2_{1:1}$ for nitrifiers, denitrifiers, PAOs, GAOs, and filamentous organisms was 32.62%, 62.65%, 53.46%, 53.31%, and 62.86%, respectively (Fig. 5a).

To further understand the prediction results of functional groups, we also analyzed the importance weights of environmental factors in their predictive models. By performing Ward clustering analysis on the importance weights of environmental factors in these predictive models of different functional groups, we found that the importance of environmental factors in predicting PAOs and GAOs was the most similar, followed by nitrifiers and denitrifiers (Fig. 5b), which implied a consistent contribution of environmental factors when predicting relevant functions. Overall, the design and operation parameters BY and Denitri, inflow condition IndConInf, physicochemical properties sludge loading (F/M), and nitrate nitrogen concentration (NO_3-N) were important for predicting nitrogen and phosphorus removal function, while climatic conditions Lat and SMT, design and operation parameter Nitri, and physicochemical properties SVI and MIT significantly affected the prediction of filamentous organisms. Additionally, SVI also had a crucial impact on the prediction of denitrifiers, which may be because some filamentous bacteria also function as denitrifiers [37]. To demonstrate the importance of these environmental factors, we only used the above 10 high-weight environmental factors to predict functional groups. The results showed that using only those ten factors allowed us to predict the abundance of major functional groups in AS systems with a predictive accuracy of $R^2_{1:1}$ ranging from 17.25% to 52.00% (Additional file 2: Figure S9).

In summary, the climatic conditions Lat and SMT, the design and operation parameters BY, Denitri, and Nitri, the inflow condition IndConInf, as well as some sample physicochemical properties (F/M, SVI, MIT, and NO_3-N) of AS systems all affect the prediction of functional groups. Controlling these critical environmental factors can help us regulate the performance of WWTPs, which will guide us to design more reasonable operating parameters according to environmental changes.

Discussion

In this study, we predicted the diversity and the structures of the microbial community, as well as the functional groups in AS systems using ANN models. We also evaluated the importance of environmental factors in the predictions.

The use of artificial neural network (ANN) models in this study increased the predictive power of complex systems of microbial communities. When ANN models were used to predict ASVs appearing in at least 10% of the samples, 60.82% of the $ASVs_{>10\%}$ had a prediction accuracy $R^2_{1:1}$ exceeding 30%. In a previous study, the multiple regression model could only explain about 15% of the variability in the genus-level taxonomy of a soil bacterial microbial community [22] and only predicted the top ten taxa of that community. Compared

with this previous study, our prediction accuracy was greatly improved, with the prediction range being increased to all ASVs appearing in at least 10% of samples, which proves the application potential of ANN models in predicting the complex systems of microbial communities. We recommend using the ANN model as a deep learning method in the prediction of complex microbial communities.

Using the Neutral Community Model (NCM) proposed by Sloan et al. [31], this study transformed migration from a vague qualitative concept into a number with biological meaning, the potential migration rate (m). Higher values of m indicate that a species is less limited by dispersal. The low migration rate of high-abundance taxa and high migration rate of low-abundance taxa in this study (Additional file 2: Figure S10) indicates that dispersal limitation has a significant effect on high-abundance taxa, but not on low-abundance taxa, which is consistent with findings for some ecosystems [38, 39]. High-abundance taxa with a low migration rate will appear in some samples due to environmental selection [40], and their relative abundance can be well predicted using these environmental factors. However, low-abundance taxa with high migration rates usually appear in a sample when the migration occurs and the spatial heterogeneity of the sample provides them with ecological niches. Neither the randomness of migration nor the spatial heterogeneity of samples was reflected in our input environmental variables, as such, these environmental factors were less predictive of low-abundance taxa. *Nitrosomonas*, the main genus of nitrifiers, is a group with a low relative abundance and high migration in the AS system (Table S2), so the predictability of nitrifiers is low (Fig. 5a). In addition, low-abundance taxa has been reported to have higher abundance variability than high-abundance taxa [30], and prediction targets with higher variability are not conducive to the stability of the predictive model, further explaining why the predictability of the relative abundance of high-abundance taxa was significantly higher than that of low-abundance taxa.

The importance of low-abundance rare species in many ecosystems has been demonstrated [27, 41]. These species play important roles in the community by providing necessary traits or acting as partners in interspecific interactions [42, 43]. To gain a better understanding of the importance of rare taxa in the microbial community, it is essential to develop a prediction model that accurately identifies low-abundance species. As the microbial community is influenced by both abiotic environmental factors [16] and species interaction [44], a machine-learning prediction model that considers the mechanism of microbial interaction may improve the prediction accuracy of rare species.

The weight of environmental factors in the predictive model reflects the influence of environmental factors on the corresponding prediction targets to a certain extent. For example, our results showed that the most important environmental factors affecting the prediction of evenness and richness were DO and IndConInf, respectively. Evenness and richness are two critical indicators to measure the diversity of ecological communities. The former describes species differences, and the latter describes the number of species. Previous studies have demonstrated that relative abundances of some functional taxa are sensitive to changes in DO [45, 46], and the abundance of these functional bacteria reflects the differences in species abundance of the community. Therefore, DO has a high weight in predicting the evenness of microbial communities in AS systems. Industrial wastewater contains many toxic and harmful substances [47, 48], which many microorganisms cannot survive. Therefore, industrial wastewater directly affects the population of microorganisms [49], and IndConInf plays an important role in predicting the richness of microbial communities in AS systems.

In addition, the impact of environmental factors on microbial taxa may be related to the specific function. The environmental factors with top weights in predictive models of nitrogen removal-related taxa ASV6 and ASV142 were AtInfTN, Nitri, and NO_3-N (Table S4). The SVI is very important for the prediction of filamentous organisms (Fig. 5b), which is because filamentous bacteria will cause sludge bulking and foaming [50], and thus affect the SVI. The Denitri has the greatest impact on PAOs and GAOs, which is consistent with the denitrification capacity of the typical PAOs genus *Ca_Accumulibacter* and GAOs genus *Ca_Competibacter* [51]. This correspondence between functions and environmental factors indicates that environmental factors with high weights in predicting microbial taxa may play an essential role in environmental filtering in the deterministic process of community assembly.

Important factors that cannot be identified using traditional methods may be highlighted by ANN modeling. Conventional studies on AS systems have only focused on the correlation relationship between environmental factors and microbial communities [16, 17, 52], which limited the scope of consideration for key environmental factors. For example, we found that whether industry wastewater source contained in inflow (IndConInf) was a significant predictor in $ASVs_{>10\%}$ predictive models in this study. This finding is consistent with earlier research which has demonstrated notable differences in microbial communities between industrial and municipal sewage [14, 53], suggesting that the IndConInf may influence the microbial community structure of AS systems [54].

However, our correlation analysis did not reveal a significant association between IndConInf and the microbial community structure (Table S5). Actually, the correlation analysis of environmental factors correlation analysis is limited in its ability to capture more complex relationships and can only reveal simple linear or monotonic associations [16, 55]. Therefore, its application in exploring the impact of environmental factors is constrained. By analyzing the importance weights of environmental factors in predictive models, this study illuminated variables that require further attention and that can better predict and control the microbial community of AS systems.

Although our work has made some contributions to the prediction and interpretation of the microbial community structure in AS systems, we still cannot explain the weights of some environmental factors in the predictive model due to the black-box characteristics of the ANN model. Our results show that environmental factors with low skewness and low kurtosis distribution are more likely to have higher weights in predicting the relative abundance of microbial taxa, which we cannot explain using current knowledge. Increasing the interpretability of the ANN model will help us better use this powerful predictive tool to analyze our concerns, which is also the future direction of machine learning-based big data analysis.

Conclusions

In this work, we used an ANN model to predict the structure of microbial communities in global AS system, including alpha diversity, ASVs appearing in at least 10% of samples, core taxa, and major functional groups. We found that taxa with high relative abundance, high occurrence frequency, and low estimated migration rate were more accurately predicted by the ANN model. Furthermore, the presence of industrial wastewater in the inflow significantly impacted the prediction of microbial communities, as demonstrated by the weight analysis of environmental factors in the ANN models. This finding implies the important role of industrial wastewater in shaping microbial communities in AS systems. Overall, our findings highlighted the importance of the ANN model in predicting the complex microbial communities. They also provide new insights into the predictability of microbial taxa and the influence of environmental factors on microbial communities.

Methods

Datasets

This study used a previously published dataset of 1186 activated sludge samples taken from 269 WWTPs in 23 countries across 6 continents. In addition to 16S rRNA sequencing data of these sludge samples, associated metadata conforming to the Genomic Standards Consortium's MIxS and Environmental Ontology Standards [56] were also provided by plant managers and investigators.

Among the 1186 samples collected in the previous study, some were from different sampling points of the same WWTP, and some were obtained from the same sampling point at different times. As such, the environmental factors and community structure between these samples may have high similarities [3] and when evaluating a model with all 1186 samples, it may overestimate the generalization ability of its predictions. Therefore, we removed the samples with the same environmental information and minimal weighted-UniFrac distance (no more than Q1-3*IQR, Q1 is the first quartile, and IQR is Inter-Quartile Range) of the microbial community in these 1186 original samples and used the remaining 777 samples (no data leakage) for subsequent construction and evaluation of the predictive model.

Data preprocessing

To ensure the accuracy, completeness, and consistency of the data, we preprocessed the original data before building the machine learning predictive model.

Metadata preprocessing

For the metadata obtained from previous studies (reference [3]), to reduce the redundancy of environmental data, we first removed some non-numerical variables of multiple categories that are difficult to operate and some variables with no practical meanings, such as site name, city name, etc. The remaining variables were used to train the model and their abbreviations and meanings are shown in Table S7. To have a clearer understanding of the environmental factors, we classified the different types of environmental factors used for prediction [3], including climate conditions, design and operation parameters, inflow conditions, effluent conditions, and physicochemical properties of samples (Table S7). Then, the *LabelEncoder* algorithm was used to numeric binary non-numeric variables, and missing values were completed according to the two-nearest neighbor principle. Additionally, all environmental factors were normalized to 1–100 to eliminate dimensional influence [24].

The final environment data for input in our machine learning predictive models is shown in Table S8.

Sequencing data preprocessing

The original microbial sequencing data were processed using Quantitative Insights Into Microbial Ecology (QIIME2) software (http://qiime2.org) [57]. All paired reads were merged, quality filtered, then denoised through the DADA2 plugin [58] to clustered into 100%

amplicon sequence variants (ASVs). Then, ASVs classified as fungi, ASVs with unassigned taxonomy at the domain level, and ASVs annotated as mitochondria or chloroplasts were removed so that only bacterial and archaeal sequences were retained. Singletons (ASVs with only one sequence) were discarded before further analysis to reduce the impact of sequencing errors. Then we rarefied each sample to 20,954 sequences, to obtain the maximal observation of both samples and features, from which 46,628 ASVs were obtained. The final feature sequences were taxonomically classified using the MiDAS4 reference database [36], and phylogenetic trees were generated using phylogeny plugins for further analysis.

Alpha diversity indices such as the Shannon–Wiener index, ASV count (species richness), and Pielou's evenness were calculated using the vegan package of R 4.0.3 software (http://www.r-project.org) according to the final feature table. Faith's phylogenetic diversity was calculated using the Picante package of R 4.0.3 software according to the feature table and phylogenetic tree. The relative abundance of each ASV was also calculated from this feature table. Together, these microbial community features served as target variables for our AS community predictive models.

Artificial neural network model

We employed the PyTorch (v1.7.1, https://pytorch.org/) library in python 3.8 to build fully connected artificial neural networks (ANNs). After testing, the three-layer network (including a hidden layer), with *relu* and *sigmoid* as activation functions between layers, had an excellent prediction effect on the condition that the network topology was relatively simple.

The first layer was the input layer, and this study's input data was the sewage treatment plants' environmental data (Table S8). Therefore, there were 48 nodes in the first layer. According to previously studied algorithm optimization results [59], the internal hidden layer had 97 nodes ($2n+1$, where n is the number of input nodes). Meanwhile, the output layer had 1 node, corresponding to the index of alpha-diversities, the relative abundance of different ASVs, or the abundance of functional groups (Fig. 2a). We built predictive models separately for each target to minimize prediction errors.

There were many random operations in the model training process, which made the results inconsistent after repeatedly running the same code. We set a fixed global seed for the random number generator to obtain repeatable training results. All models were trained twenty times by different seeds to avoid the deviation caused by each randomization, and the averaged results were used for further analysis.

Alpha diversity and microbial taxa abundance predictive model

For alpha diversity, we established predictive models for the Shannon–Wiener index, Pielou's evenness index, species richness, and Faith's phylogenetic diversity. For taxa, the relative abundance of taxa with low occurrence frequency was zero in many samples, which made it difficult for the model to learn useful information on the training set (underfitting). Therefore, only the relative abundance of ASVs present in at least 10% of samples (named $ASVs_{>10\%}$, corresponding $ASVs_{<10\%}$ represent ASVs that appear in no more than 10% samples) were modeled to predict.

There were 777 samples to build the alpha diversity or $ASVs_{>10\%}$ abundance predictive models. To reduce the risk of overfitting during hyperparameter optimization, we performed fourfold cross-validation in the training process. As a result, these models were developed by applying fourfold cross-validation to 80% of the total samples. Test sets comprising the remaining 20% of the whole samples were used to evaluate the performance of the models. All models were trained 20 times by different seeds to avoid obtaining model bias. Finally, the model performance was assessed based on the averaged results.

In the training processes of ANN models, the coefficient of determination (R^2) and mean square error (MSE) were used to evaluate the accuracies and losses. After optimization of hyperparameters, we used an *Adam* optimizer with a batch size of 256, a learning rate of 0.00001, a drop-out of 0, and a weight decay of 0.01 to train these models. To obtain the number of iterations when the model was optimal, we repeatedly tested the variation of R^2 and MSE with the number of iterations (Additional file 2: Figure S11). The results showed that after reaching 5000 iterations, the R^2 and MSE of most models started to remain stable. The number of iterations for all models was set to 10,000, considering the trade-off between the time cost of iteration and the lowest losses.

From neutral community model to neutral and non-neutral partitions

To determine the potential importance of stochastic processes on WWTP community assembly, we used a neutral community model (NCM) to predict the relationship between an ASVs' occurrence frequency and their relative abundance across the wider metacommunity [31]. The model is an adaptation of the neutral theory adjusted to large microbial populations. In this model, m is an estimate of dispersal between communities, being the estimated migration rate. Because the estimation of migration rate m is affected by the number of sequences in samples, we flattened the number

of sequences in each sample to 2000 before fitting the neutral community model, allowing us to compare estimated migration rates for different microbial partitions.

In this study, all 46,628 ASVs were separated into three partitions depending on whether they occurred more frequently than (above partition), less frequently than (below partition), or within (neutral partition) the 95% confidence interval of the NCM predictions [60]. To explore the effect of the potential migration rate of ASVs on their predictability, we analyzed and compared the predictive accuracy of different (above, neutral, and below) partitions belonging to the common $ASVs_{>10\%}$ sub-community.

Definition of core taxa

A global-scale core microbial subcommunity of WWTP was determined based on multiple reported measures. In this report, we explored the predictability of microbial taxa at the ASV level (100% similarity), as such the classification criteria for core ASVs were slightly different than those for core OTUs [3]. First, 'overall abundant ASVs' were filtered out according to the mean relative abundance (MRA) across all samples. We selected all top 1% ASVs as overall abundant ASVs. Second, 'ubiquitous ASVs' were defined as ASVs with an occurrence frequency in more than 20% of all samples. Finally, 'frequently abundant ASVs' were selected based on their relative abundances within a sample. In each sample, the ASVs were defined as abundant when they had a higher relative abundance than other ASVs and made up the top 80% of the reads in the sample [34]. A frequently abundant ASV was defined as abundant in at least 10% of the samples. Following the same criteria described above, a core ASV should be one that was from the top 1% ASVs, a core ASV also had to be detected in more than 20% of the samples and dominant for more than 10% of the samples. Corresponding to the core taxa, ASVs that did not meet the above three conditions were called non-core taxa.

Statistical analysis

All alpha diversity measures were conducted using the vegan and Picante packages of R (v. 4.0.3). Unless indicated otherwise, an unpaired, two-tailed, two-sample Student's *t*-test was performed for comparative statistics using the *t.test* function in the stats package of R 4.0.3. Linear correlation analyses between different parameters were implemented using the *lm* function in the stats package of R 4.0.3. All analysis and graphing were done using R4.0.3 or python 3.8.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-023-01519-9.

Additional file 1. Supplementary analysis. Grouping and Comparisonof $ASVs_{>10\%}$.

Additional file 2: Figure S1. Ranking of importance weights ofenvironmental factors in different alpha-diversities predictive models. **FigureS2.** a. Comparison ofintra- and inter-group Bray-Curtis similarity between predicted and observedcommunities. b. Average prediction accuracy $R^2_{1:1}$of microbial taxa at different taxonomic levels. **Figure S3.** Environmental factor importance weights andPearson's correlation coefficients. **FigureS4.** Correlation ofcorrelation coefficients of environment factors with $ASVs_{>10\%}$subcommunity, skewness, and kurtosis of normalized environment variables withtheir Garson's connection weights. **Figure S5.** a. Comparison of predictiveaccuracy $R^2_{1:1}$ between low,medium, and high abundance taxa. b. Comparison of predictiveaccuracy $R^2_{1:1}$ between low,medium, and high-frequency taxa. c. Correlation of relative abundance with the occurrencefrequency of ASVs. d. Correlation of the $R^2_{1:1}$in test sets with the coefficient of variation of ASVs. **Figure S6.** Comparison ofaverage relative abundance and occurrencefrequency between above, neutral, and below partitions. **Figure S7.** Fit of theneutral community model (NCM) of above, neutral, and below partitions. **Figure S8.**The taxonomic composition, average relative abundance, occurrence frequency,and estimated migration rate of core and non-core taxa. **Figure S9.** Predictionof functional groups with 10 high-weight environmental factors. **Figure S10.** Fitof the neutral community model (NCM) of high abundance, medium abundance, andlow abundance subcommunities. **Figure S11.** Changes of mean square errors (MSE) andcoefficients of determination (R^2) on the validation set with epochswhen training the model.

Additional file 3: Table S1. Alpha-diversitiesof AS system. **Table S2.** Summary of ASVs belonging $ASVs_{>10\%}$ sub-community.**Table S3.** Summary of microbial taxa at different taxonomic levels. **Table S4.** Averageimportance weights of environmental factors in different ASVs predictivemodels. **Table S5.** Summary of different environment variables. **Table S6.** Summaryof genera belonging to major functional groups. **TableS7.** Abbreviations, meanings, and types of environment variables. **Table S8.** Numericaland normalized environmental data.

Acknowledgements

The authors thank the Global Water Microbiome Consortium (GWMC) and all the people involved for providing samples and plant metadata. We thank the High-performance Computing Platform of Peking University for providing the computing platform.

Authors' contributions

X.L. conceived the study and performed all analysis and computation. X.L. and Y.N. co-wrote the paper, and X.L.W. revised it. All authors discussed the results and commented on the article. The author(s) read and approved the final manuscript.

Funding

This study has received funding from the National Key R&D Program of China (2018YFA0902100 and 2021YFA0910300) and the National Natural Science Foundation of China (91951204, 32130004, 32161133023, and 32170113).

Availability of data and materials

The raw data in this study is from reference [3]. All analyzed data in this study is available in Additional file 3. The source code is available at https://github.com/Neina-0830/WWTP_community_prediction.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 September 2022 Accepted: 16 March 2023
Published online: 28 April 2023

References

1. Jones ER, van Vliet MTH, Qadir M, Bierkens MFP. Country-level and gridded estimates of wastewater production, collection, treatment and reuse. Earth System Sci Data. 2021;13:237–54.
2. van Loosdrecht MCM, Brdjanovic D. Anticipating the next century of wastewater treatment. Science. 2014;344:1452–3.
3. Wu L, Ning D, Zhang B, Li Y, Zhang P, Shan X, Zhang Q, Brown MR, Li Z, Van Nostrand JD, et al. Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nat Microbiol. 2019;4:1183–95.
4. Fang H, Cai L, Yu Y, Zhang T. Metagenomic analysis reveals the prevalence of biodegradation genes for organic pollutants in activated sludge. Bioresour Technol. 2013;129:209–18.
5. Yu X, Nishimura F, Hidaka T. Effects of microbial activity on perfluorinated carboxylic acids (PFCAs) generation during aerobic biotransformation of fluorotelomer alcohols in activated sludge. Sci Total Environ. 2018;610–611:776–85.
6. Wang M, Chen H, Liu S, Xiao L. Removal of pathogen and antibiotic resistance genes from waste activated sludge by different pre-treatment approaches. Sci Total Environ. 2021;763:143014–23.
7. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, Albertsen M. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. Nat Commun. 2021;12:2009–21.
8. Nierychlo M, Andersen KS, Xu Y, Green N, Jiang C, Albertsen M, Dueholm MS, Nielsen PH. MiDAS 3: An ecosystem-specific reference database, taxonomy and knowledge platform for activated sludge and anaerobic digesters reveals species-level microbiome composition of activated sludge. Water Res. 2020;182:115955–66.
9. Herold M, Martinez Arbas S, Narayanasamy S, Sheik AR, Kleine-Borgmann LAK, Lebrun LA, Kunath BJ, Roume H, Bessarab I, Williams RBH, et al. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. Nature Commun. 2020;11:5281–94.
10. Griffin JS, Wells GF. Regional synchrony in full-scale activated sludge bioreactors due to deterministic microbial community assembly. ISME J. 2017;11:500–11.
11. Nielsen PH, Saunders AM, Hansen AA, Larsen P, Nielsen JL. Microbial communities involved in enhanced biological phosphorus removal from wastewater-a model system in environmental biotechnology. Curr Opin Biotechnol. 2012;23:452–9.
12. Zhang T, Shao MF, Ye L. 454 pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. ISME J. 2012;6:1137–47.
13. Shchegolkova NM, Krasnov GS, Belova AA, Dmitriev AA, Kharitonov SL, Klimina KM, Melnikova NV, Kudryavtseva AV. Microbial Community Structure of Activated Sludge in Treatment Plants with Different Wastewater Compositions. Front Microbiol. 2016;7:90–104.
14. Ibarbalz FM, Figuerola EL, Erijman L. Industrial activated sludge exhibit unique bacterial community composition at high taxonomic ranks. Water Res. 2013;47:3854–64.
15. Zhang B, Ning D, Yang Y, Van Nostrand JD, Zhou J, Wen X. Biodegradability of wastewater determines microbial assembly mechanisms in full-scale wastewater treatment plants. Water Res. 2020;169:115276–84.
16. Wei Z, Liu Y, Feng K, Li S, Wang S, Jin D, Zhang Y, Chen H, Yin H, Xu M, Deng Y. The divergence between fungal and bacterial communities in seasonal and spatial variations of wastewater treatment plants. Sci Total Environ. 2018;628–629:969–78.
17. Tian L, Wang L. A meta-analysis of microbial community structures and associated metabolic potential of municipal wastewater treatment plants in global scope. Environment Poll. 2020;263:114598–608.
18. Gonzalez-Cabaleiro R, Ofiteru ID, Lema JM, Rodriguez J. Microbial catabolic activities are naturally selected by metabolic energy harvest rate. ISME J. 2015;9:2630–41.
19. Bairey E, Kelsic ED, Kishony R. High-order species interactions shape ecosystem diversity. Nat Commun. 2016;7:12285–91.
20. Wang M, Liu X, Nie Y, Wu XL. Selfishness driving reductive evolution shapes interdependent patterns in spatially structured microbial communities. ISME J. 2020;15:1387–401.
21. Liu X, Wang M, Nie Y, Wu X-L. Successful microbial colonization of space in a more dispersed manner. ISME Commun. 2021;1:68–77.
22. Averill C, Werbin ZR, Atherton KF, Bhatnagar JM, Dietze MC. Soil microbiome predictability increases with spatial and taxonomic scale. Nat Ecol Evol. 2021;5:747–56.
23. Krogh A. What are artificial neural networks? Nat Biotechnol. 2008;26:195–7.
24. Larsen PE, Field D, Gilbert JA. Predicting bacterial community assemblages using an artificial neural network approach. Nat Methods. 2012;9:621–5.
25. Kuang J, Huang L, He Z, Chen L, Hua Z, Jia P, Li S, Liu J, Li J, Zhou J, Shu W. Predicting taxonomic and functional structure of microbial communities in acid mine drainage. ISME J. 2016;10:1527–39.
26. Coutinho FH, Thompson CC, Cabral AS, Paranhos R, Dutilh BE, Thompson FL. Modelling the influence of environmental parameters over marine planktonic microbial communities using artificial neural networks. Sci Total Environ. 2019;677:205–14.
27. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol. 2015;13:217–29.
28. Chapin Iii FS, Zavaleta ES, Eviner VT, Naylor RL, Vitousek PM, Reynolds HL, Hooper DU, Lavorel S, Sala OE, Hobbie SE, et al. Consequences of changing biodiversity. Nature. 2000;405:234–42.
29. Henríquez PA, Ruz GA. A non-iterative method for pruning hidden neurons in neural networks with random weights. Appl Soft Comput. 2018;70:1109–21.
30. Kim TS, Jeong JY, Wells GF, Park HD. General and rare bacterial taxa demonstrating different temporal dynamic patterns in an activated sludge bioreactor. Appl Environ Microbiol. 2013;97:1755–65.
31. Sloan WT, Lunn M, Woodcock S, Head IM, Nee S, Curtis TP. Quantifying the roles of immigration and chance in shaping prokaryote community structure. Environ Microbiol. 2006;8:732–40.
32. Song C, Fukami T, Saavedra S. Untangling the complexity of priority effects in multispecies communities. Ecol Lett. 2021;24:2301–13.
33. Pagaling E, Strathdee F, Spears BM, Cates ME, Allen RJ, Free A. Community history affects the predictability of microbial ecosystem development. ISME J. 2014;8:19–30.
34. Saunders AM, Albertsen M, Vollertsen J, Nielsen PH. The activated sludge ecosystem contains a core community of abundant organisms. ISME J. 2016;10:11–20.
35. Matar GK, Bagchi S, Zhang K, Oerther DB, Saikaly PE. Membrane biofilm communities in full-scale membrane bioreactors are not randomly assembled and consist of a core microbiome. Water Res. 2017;123:124–33.
36. Dueholm MKD, Nierychlo M, Andersen KS, Rudkjobing V, Knutsson S. Mi DASGC, Albertsen M, Nielsen PH: MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. Nat Commun. 2022;13:1908–22.
37. Nielsen PH, de Muro MA, Nielsen JL: Studies on the in situ physiology of Thiothrix spp. present in activated sludge. Environ Microbiol. 2000, 2:389–398.
38. Jiao S, Lu Y. Soil pH and temperature regulate assembly processes of abundant and rare bacterial communities in agricultural ecosystems. Environ Microbiol. 2020;22:1052–65.
39. Wu W, Logares R, Huang B, Hsieh CH. Abundant and rare picoeukaryotic sub-communities present contrasting patterns in the epipelagic waters of marginal seas in the northwestern Pacific Ocean. Environ Microbiol. 2017;19:287–300.

40. Shao Q, Sun D, Fang C, Feng Y, Wang C. Biodiversity and Biogeography of Abundant and Rare Microbial Assemblages in the Western Subtropical Pacific Ocean. Front Microbiol. 2022;13:839562–75.
41. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, Kusel K, Rillig MC, Rivett DW, Salles JF, et al. Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J. 2017;11:853–62.
42. Shade A, Jones SE, Caporaso JG, Handelsman J, Knight R, Fierer N, Gilbert JA, Dubilier N: Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. mBio. 2014, 5:e01371–01314.
43. Fetzer I, Johst K, Schäwe R, Banitz T, Harms H, Chatzinotas A. The extent of functional redundancy changes as species' roles shift in different environments. Proc Nat Acad Sci U S A. 2015;112:14888–93.
44. Ju F, Xia Y, Guo F, Wang Z, Zhang T. Taxonomic relatedness shapes bacterial assembly in activated sludge of globally distributed wastewater treatment plants. Environ Microbiol. 2014;16:2421–32.
45. Cao Y, Zhang C, Rong H, Zheng G, Zhao L. The effect of dissolved oxygen concentration (DO) on oxygen diffusion and bacterial community structure in moving bed sequencing batch reactor (MBSBR). Water Res. 2017;108:86–94.
46. Laureni M, Weissbrodt DG, Villez K, Robin O, de Jonge N, Rosenthal A, Wells G, Nielsen JL, Morgenroth E, Joss A. Biomass segregation between biofilm and flocs improves the control of nitrite-oxidizing bacteria in mainstream partial nitritation and anammox processes. Water Res. 2019;154:104–16.
47. Liang S-x. Wang J, Qin Z, Zhao C, Jin X, Chen J: Biotoxicity and by-product identification of dye wastewaters. Water Pract Technol. 2019;14:449–56.
48. Hubadillah SK, Othman MHD, Tai ZS, Jamalludin MR, Yusuf NK, Ahmad A, Rahman MA, Jaafar J, Kadir SHSA, Harun Z. Novel hydroxyapatite-based bio-ceramic hollow fiber membrane derived from waste cow bone for textile wastewater treatment. Chem Eng J. 2020;379:122396–407.
49. Yang Y, Wang L, Xiang F, Zhao L, Qiao Z. Activated sludge microbial community and treatment performance of wastewater treatment plants in industrial and municipal zones. Int J Environ Res Public Health. 2020;17:436–50.
50. Guo F, Zhang T. Profiling bulking and foaming bacteria in activated sludge by high throughput sequencing. Water Res. 2012;46:2772–82.
51. Lemaire R, Yuan Z, Blackall LL, Crocetti GR: Microbial distribution of Accumulibacter spp. and Competibacter spp. in aerobic granules from a lab-scale biological nutrient removal system. Environ Microbiol. 2008, 10:354–363.
52. Zhang J, Liu GH, Wei Q, Liu S, Shao Y, Zhang J, Qi L, Wang H. Regional discrepancy of microbial community structure in activated sludge system from Chinese WWTPs based on high-throughput 16S rDNA sequencing. Sci Total Environ. 2021;818:151751–8.
53. Fan XY, Gao JF, Pan KL, Li DC, Dai HH, Li X. Functional genera, potential pathogens and predicted antibiotic resistance genes in 16 full-scale wastewater treatment plants treating different types of wastewater. Bioresource Technol. 2018;268:97–106.
54. Sun H, Chang H, Tang W, Zhang X, Yang H, Zhang F, Zhang Y. Effects of influent immigration and environmental factors on bacterial assembly of activated sludge microbial communities. Environ Res. 2022;205:112426–35.
55. Guo H, Nasir M, Lv J, Dai Y, Gao J. Understanding the variation of microbial community in heavy metals contaminated soil using high throughput sequencing. Ecotoxicol Environ Safety. 2017;144:300–6.
56. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29:415–20.
57. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–7.
58. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3.
59. Stathakis D. How many hidden layers and nodes? Int J Remote Sensing. 2009;30:2133–47.
60. Chen W, Ren K, Isabwe A, Chen H, Liu M, Yang J. Stochastic processes shape microeukaryotic community assembly in a subtropical river across wet and dry seasons. Microbiome. 2019;7:138–53.

Publisher's Note