Modeling of a Mainstream Partial Nitrification/Anammox Process through a Hybrid Theoretical-Machine Learning Approach

Valeria Alvarado[1†], Lebing Ying[1†], Vahid Asghari[1], Shu-Chien Hsu[1*], and Po-Heng Lee[2]

[1] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong S.A.R

[2] Department of Civil and Environmental Engineering, Imperial College London, South Kensington SW7 2AZ, London, United Kingdom

†These authors contributed equally to this work and were designated as co-first authors.

* Corresponding Author: mark.hsu@polyu.edu.hk, Tel: +852-27666057

Abstract: Model simulations are vital in optimizing and predicting the performance of biological sewage treatment, especially for processes involving slow-growing bacteria. However, data records often include missing, invalid, or infrequent measurements of parameters, compromising prediction accuracy. This study used a hybrid theoretical-machine learning approach to address these issues. By leveraging stoichiometry and kinetics, missing values were calculated in limited data sets, which were then analyzed through machine learning algorithms to reveal hidden microbial interactions. The model was validated with data from a pilot-scale partial nitritation/anammox fluidized bed membrane bioreactor (PN/A FMBR) with saline sewage. The model demonstrated strong prediction performance, with random forest outperforming other algorithms, with correlation coefficients of 0.89, 0.72, and 0.80 for ammonium, nitrite, and nitrate data sets, respectively, when compared to actual values. Training sets containing 73 to 88 same-day values reached acceptable predicting performance. The results also revealed that microbial synergy in nitrogen transformation, particularly in the partial denitrification from nitrate to nitrite linked to Anammox in responding to a low DO supply, was evident in this PN/A FMBR. Additionally, key parameters, including temperature, pH, and specific microbiomes, were identified as critical for predicting PN/AFMBR performance, highlighting significant microbial interactions that warrant further investigation.

Keywords: Wastewater treatment, Anammox, Theoretical-machine learning, Partial nitritation, Microbial interactions, Fluidized bed membrane bioreactor

Synopsis: This study presents an innovative hybrid Theoretical-Machine Learning model for sewage treatment, calculating missing values and revealing crucial microbial interactions to enhance effluent prediction accuracy.

1. INTRODUCTION

Traditional biological nitrogen removal (BNR) methods in sewage treatment plants (STPs), such as nitrification/denitrification (N/D) are energy-intensive and produce high levels of sludge and green house gas. Alternative approached like partial nitrification (PN) and anaerobic ammonium oxidation (anammox) offer more sustainable solutions.[1] PN nitrifies half the ammonium (NH_4^+) in the influent to nitrite (NO_2^-) by ammonium oxidizing bacteria, while anammox removes the remaining NH_4^+ with NO_2^- to produce nitrogen gas (N_2) by anammox bacteria [2–7]. Coupled PN/anammox (PN/A) process produces N_2 and small amounts of nitrate (NO_3^-). Integration of the PN/A process in a membrane bioreactor (MBR) has shown high nitrogen removal efficiency (~89.5%) has been theoretically demonstrated [3,8]. In addition to the benefits of the PN/A processes, its integration to an MBR results in high-quality effluent [9–11] at short hydraulic retention times (HRTs) and high loading rates.[2,12]

Research indicates that the PN/A process is complicated and highly susceptible to environmental factors, including temperature, pH, dissolved oxygen, the presence of heavy metals and the concentration of ammonium and nitrite [13]. Additionally, the sensitivity of anammox bacteria and the unstable suppression of nitrite-oxidizing bacteria (NOB) in the complex mainstream environment vastly limited anammox applications.[14,15] This sensitivity necessitates precise control of operational parameters to avoid costly and time-consuming recovery from disturbances..[16–18] Modelling has proven to be an invaluable tool for enhancing understanding of recent anammox-based processes and helping in the simulating various operational strategies [19–21]. In the context of PN/A systems, the primary focus was on autotrophic nitrogen removal carried out by ammonia-oxidizing bacteria (AOB), NOB and anammox bacteria. However, recent research suggests that heterotrophic microorganisms may also play a critical role in these systems.[20] While autotrophic processes dominate the PN/A pathway, heterotrophs can contribute by influencing nitrite availability and competing for

resources, thereby affecting overall system efficiency. Therefore, more efforts are required to characterize the microbial interactions not only between AOB and NOB but also between autotrophic and heterotrophic organisms to improve nitrogen removal efficiency.

Understanding these microbial synergies, alongside the multiple pathways contributing to nitrogen transformation, is crucial for optimizing PN/A performance. Therefore, further research is necessary to better characterize microbial interactions by identifying multiple pathway contributions, influencing factors and energy-efficient operational strategies for the successful implementation of these processes. [19,22]

Models which include microbial community diversity enable the understanding of microbial community shifts and composition link to control strategies, stability, and performance.[23] To address the complexities of modeling such systems, tools like the activated sludge model (ASM) series[24,25] and anaerobic digestion model 1,[26] are commonly used, integrated with software such as AQUASIM[27], GPS-X[28], and Biowin [29] with special modules for MBR simulations [30,31]. PN/A process in granular sludge reactors have been studied with AQUASIM's multi-substrate [17,32–35] and multi-species biofilm models. These models are invaluable for predicting system behavior and optimizing operational strategies[36] but often suffer from a lack of standardized methods for parameter calculation and challenges in scaling up from biofilm-based models to real-world applications.[32,37]

In recent years, the statistical-empirical approach through machine learning algorithms have been applied in the sewage treatment field for sewage quality influent and effluent prediction[38–40], soft measurement [41] and automatic control of STPs [42,43] due to its ability for handling large datasets and dealing with complex, nonlinear relationships and accurate predicting. Furthermore, recent advances in interpretable techniques,[44,45] now offer insights into the key factors driving biological sewage treatment. In practical wastewater treatment scenarios, data gaps are common,[37,46] particularly with key operational parameters like microbial community

abundances and real-time nitrogen concentrations.[45,47] These gaps can significantly hinder the accuracy of outcome predictions. Additionally, relying solely on machine learning can fail to capture the complex biological interactions within these systems, often resulting in incomplete process understanding.[48] While there have been many models developed for PN/A process[19,86], few theoretical- or statistical-empirical modelling approaches have specifically addressed the PN/A process in an anoxic MBR with granulated activated carbon (GAC) and a fluidized-bed with sewage recirculation. Moreover, the influence of microbial community abundance on nitrogen removal in this distinctive setup has seldom been analyzed through theoretical or statistical-empirical modeling approach. In this context, this paper introduces a novel hybrid theoretical-statistical approach designed specifically for data-limited settings. This approach is applied to a partial nitrification/anammox fluidized-bed membrane bioreactor (PN/AFMBR) to predict effluent nitrogen compounds effectively. It also aims to identify crucial parameters influencing biological nitrogen removal (BNR) by incorporating operational, water quality, and microbial data. By calculating missing values and revealing hidden microbial interactions that traditional methods overlook, this study successfully bridges gaps in optimizing the performance of biological sewage treatment systems, particularly those relying on slow-growing bacteria.

2. MATERIALS AND METHODS

2.1 PN/AFMBR description

The pilot-scale PN/AFMBR for mainstream treatment of settled sewage obtained from chemically enhanced primary treatment was installed at the Stonecutters Island sewage treatment works (STWs) in Hong Kong, as shown in Figure S1 [49]. Sewage in Hong Kong contains high salinity because about 80% of the population is served with treated seawater for toilet flushing. The reactor operated for 523 days with final working volume of 1.23 m^3 [4] The

acclimation period was about 147 days when the operation was shifted from batch to continuous mode. Settled GAC occupied 56% of the reactor, while fluidized GAC occupied 100% of the reactor height. The PN/AFMBR contained five submerged hollow-fiber membrane (pore size: 0.1 μm) modules made of polyvinylidene fluoride with a total surface area of 55 m^2. The upflow velocity was 69.2 m/h. The reactor was inoculated with seed from an anaerobic digester in Tai Po STWs.

The operation of the PNAFMBR was automated with an integrated programmable logic controller (PLC) unit including sensors for temperature, dissolved oxygen (DO), and pH. Transmembrane pressure (TMP) was recorded with a pressure gauge installed in the effluent line. The average flow rate was adjusted according to three different HRTs (i.e., 15,10 and 8 hr).[4] Chemical oxygen demand (COD), ammonium, nitrite, and nitrate concentrations were determined according to standard methods.[50]

DNA extraction of biomass samples (i.e., seed sludge, suspension sludge in the PN/AFMBR, and biofilm from GAC in the PN/AFMBR) was performed with PowerSoil DNA extraction kit (MoBio Laboratories, Carlsbad, CA, USA). 16S rRNA gene was sequenced with Illumina Mi-Seq platform (PE250) using primers 515F and 806R. BlastKOALA was utilized for genome functions characterization and functional pathways reconstruction [4,51]. Details of the pilot-scale PN/AFMBR, its operating conditions and microbial community results are reported in Huang et al [4].

The methodology, as described in **Figure 1**, starts with analyzing data for same-day information. Missing values are then calculated theoretically. Data sets are separated in two groups depending on the number of input process parameters. Group 1 includes operation parameters and influent nitrogen compounds concentrations. Besides the input process parameters in group 1, microbial community abundances are included in Group 2. The machine learning algorithms are then developed for each data set in Group 1. The best machine learning

algorithm is identified through a ranking approach. Analysis of different data set sizes – including Group 2 – and feature selection are conducted with the machine learning algorithm with the best performance.

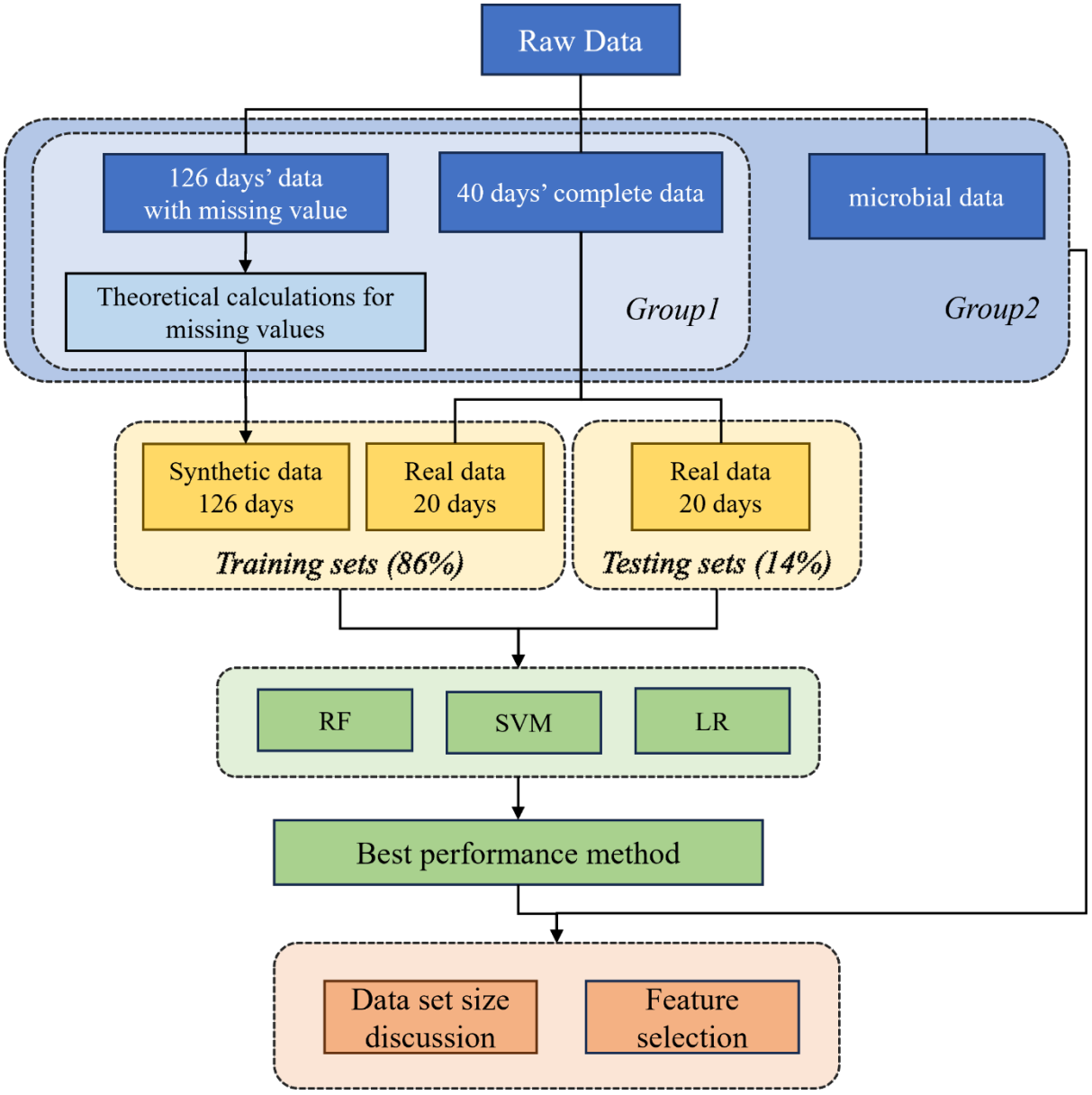


Figure 1. Overall framework of the hybrid theoretical-statistical approach, RF: random forest, and SVM: support vector machine, LR: linear regression.

2.2 Dataset preparation

2.2.1 Operational data description

To build the datasets, data was screened for same-day information on six operational parameters (i.e., HRT, temperature, transmembrane pressure, average flow rate, DO, pH), three influent and three effluent nitrogen compounds concentrations (i.e., ammonium, nitrite, and nitrate), and 56 microbial community abundance percentages in the suspension and biofilm samples in the PN/AFMBR (e.g., *Nitrosomonas*, *Nitrospira*, and *Candidatus* Kuenenia). The available data for the analysis was collected for 166 days during the period of April 2018 to October 2018 (day 181 to 346).

Excluding temperature, all operational parameters were controlled with the PLC unit. Temperature, TMP and pH in the reactor ranged from 25 to 35ºC, -0.1 to -0.01 kg/cm^2 and 7.2 to 7.8, respectively. pH was increased with sodium carbonate solution when the value decreased below 7.2. Even if DO set point was 0.1 mg/L, the average DO value from the measurements of five DO sensors installed along the reactor were recorded. Operational parameters measurements were frequently recorded (3 to 6 times per week) from the PLC unit. Concentrations of inorganic nitrogen compounds were routinely measured every 3 to 5 days at the influent and effluent of the PN/AFMBR. The nitrogen compounds influent concentrations were considered as uncontrollable variables. Samples for microbial community analysis were collected 5 times in the total period of data collection on Days 0, 216, 259, 308, and 332.[4] The explanation of dates and frequency of microbial community analysis sampling is explained in **Supporting Method S2**. At genus level, 28 microbiomes were identified for each of the seed sludge sample (S0), suspension sludge samples (S1 to S4), and biofilm samples from the GAC (B1 to B4). Microbial community abundance percentages are assumed as controllable because operation parameters might inhibit the growth of a group of microbiomes. For example, limited DO (< 0.1 mg/L) is expected to suppress nitrite oxidizing bacteria. Yet, the microbial abundance of different microbiomes in the seed sludge was out of the control of this study.

Given the disparity on the frequency of data collection, only one day (Day 332) has same-day information for operational parameters, influent and effluent nitrogen concentrations, and microbial community abundance. From Day 181 to 346, data for operational parameters were collected for 105 days, whereas data for influent and effluent nitrogen compounds were measured for 50 days. Data for microbial community abundance was obtained for 5 days. Records with missing values have been eliminated[41,52] or calculated from previous data[53] in other models based on machine learning methods. In this study, missing data for input and output process variables were calculated under theoretical assumptions to provide more professional insights.

Table 1 Statistical description of input and output process variables in data sets of Groups 1 and 2 according to Huang et al. [4]

Group 1	*Group 2*	*Process variables*	*Input*	*Min.*	*Max.*	*Mean*	*Std. Dev.*
✓	✓	Operational parameters	HRT (h)	8.000	15.000	12.663	2.881
			Temperature (ºC)	24.810	34.46	30.377	1.890
			TMP (kg/cm^2)	-0.100	-0.010	-0.048	0.016
			Flow rate (m^3/h)	0.020	0.940	0.109	0.093
			Dissolved oxygen (mg/L)	0.030	0.330	0.090	0.042
			pH	7.160	7.770	7.528	0.070
✓	✓	Influent (mg/L)	Ammonium	11.170	44.000	22.040	5.083
			Nitrite	0.000	1.270	0.025	0.120
			Nitrate	0.000	7.130	2.831	1.629
	✓	Microbial data	Microbial community abundance (%):	*	*	*	*
1	2		Output				
✓	✓	Effluent & influent difference(mg/L)	Ammonium	5.272	43.910	18.809	5.589
			Nitrite	-9.779	1.230	-1.713	2.222
			Nitrate	-14.010	5.470	-1.440	4.278

*Notes: *The statistical description of the microbial community abundance includes 28 genera for suspensions in the PN/AFMBR and 28 genera for biofilm in the GAC of the PN/AFMBR are listed in* ***Table S8–S9****.*

Input process variables include operational parameters, nitrogen compounds concentrations in the influent and microbial community abundance percentages in the suspension and biofilm within the PN/AFMBR (**Table 1**). Missing data for TMP and DO were assumed as the average between the previous and following day but circumstances on-site were also considered.[54] For example, TMP was assumed lower than average during days 239 to 241 due to power shortcuts

on-site. On Day 256, the DO was set to 0 mg/L thus the average DO in the reactors was assumed as 0.04 mg/L as measured for Day 255. Interpolation was used to assume the missing data for flow rate and pH.[55] In addition, the on-site observations were considered. For example, flow rate on Day 276 was below the measurement range, thus 0.055 m^3/hr was assumed. The missing data for temperature were assumed by the average of the previous and the following day. If temperature data for two days were missing, then the missing values were interpolated from one day before and one day after. Missing values for influent ammonium concentration were assumed as a random value within a range obtained from measurements on April 2018 to October 2018[4] and April 2016 to October 2016.[56] Nitrite concentrations in the influent were assumed as 0 mg/L for all the missing data because all the measurements in this period were below 0.6 mg/L, except for day 274 (1.27 mg/L). Missing data for nitrate concentration in the influent were assumed as a random number within the range of the measured maximum and minimum concentrations for each month (April to October 2018). Theoretical calculations based on stoichiometry and microbial kinetics have been previously determined to describe the biochemical reactions occurring in the pilot-scale PN/AFMBR through a thermodynamics electron equivalent (TEE) model.[22] Even though microbial growth can be established through stoichiometry and kinetics, the microbial community abundance are only limited within known microbiomes. Given that quantification of total active biomass is required to calculate the percentage of microbial community abundance for each genus of known microbiomes, missing percentages were assumed with linear or exponential regression between the 5 days measured.

Output process variables were calculated as the difference between effluent and influent nitrogen compounds concentrations (**Table 1**). Effluent nitrogen compounds concentrations were measured for 50 days; thus, the TEE model developed by Alvarado[22] was used to validate the date of 50 measurements from Huang et al.[4]

2.2.2 Theoretical calculation of missing data

Missing values for effluent concentrations were determined based on the TEE model. The TEE model assumes that biological reactions are based on substrate partitioning and cellular framework, quantifies microbial community abundances in various sludge types, and includes multiple biological processes (e.g., organic oxidation, denitrification, sulfate reduction) with iterative calculations for influent and effluent nitrogen compounds, utilizing design criteria from existing literature and experimental data to estimate removal efficiencies and account for missing data using specified thresholds and regression methods. The detailed calculation processes and its correlated assumptions are listed in **Supporting Method S1**. **Table S1** lists typical design criteria for the biological processes in the PN/AFMBR included in the TEE model. Ammonium concentration (in mg/L) in the effluent (${NH_4^+}_{eff}$) is calculated by including ammonium concentration in the influent (${NH_4^+}_{inf}$); ammonium produced during organic oxidation (${NH_4^+}_{O.O,p}$), fermentation (${NH_4^+}_{Fer.,p}$), heterotrophic denitrification with nitrite ($+{NH_4^+}_{HD,NO_2^-,p}$), heterotrophic partial denitrification with nitrate reduction to nitrite (${NH_4^+}_{HPD,NO_3^-,p}$), dissimilatory nitrate reduction to ammonium with sulfur reduction (${NH_4^+}_{DNRA,p}$), sulfate reduction (${NH_4^+}_{SR,p}$), and heterotrophic denitrification with nitrate (${NH_4^+}_{HD,NO_3^-,p}$); and, ammonium consumed during aerobic ammonium oxidation (${NH_4^+}_{AAO,c}$), nitrite oxidation (${NH_4^+}_{N.O,c}$), autotrophic partial denitrification with nitrate reduction to nitric oxide (${NH_4^+}_{APD,c}$), anammox (${NH_4^+}_{Anx,c}$), autotrophic denitrification with nitrate (${NH_4^+}_{AD,NO_3^-,c}$), and methanol oxidation (${NH_4^+}_{MO,c}$) as shown in Eq. 1. Nitrite concentration (in mg/L) in the effluent (${NO_2^-}_{eff}$) is determined by including nitrite concentration in the influent (${NO_2^-}_{inf}$); nitrite produced in aerobic ammonium oxidation (${NO_2^-}_{AAO,p}$), and heterotrophic partial denitrification with nitrate reduction to nitrite (${NO_2^-}_{HPD,NO_3^-,p}$); and, consumed in heterotrophic denitrification with nitrite (${NO_2^-}_{HD,NO_2^-,c}$), nitrite oxidation (${NO_2^-}_{N.O.,c}$), and anammox

($NO_{2\ Anx.,c}^{-}$) as shown in Eq.2. Nitrate concentration (in mg/L) in the effluent ($NO_{3\ eff}^{-}$) is calculated by including nitrate concentration in the influent ($NO_{3\ inf}^{-}$); nitrate produced during nitrite oxidation ($NO_{3\ N.O.,p}^{-}$), and anammox ($NO_{3\ Anx.,p}^{-}$); and, nitrate consumed in autotrophic partial denitrification with nitrate reduction to nitric oxide ($NO_{3\ APD.,c}^{-}$), autotrophic denitrification with nitrate ($NO_{3\ AD.,NO_3^-,c}^{-}$), dissimilatory nitrate reduction to ammonium with sulfur reduction ($NO_{3\ DNRA,c}^{-}$), heterotrophic partial denitrification with nitrate reduction to nitrite ($NO_{3\ HPD,NO_3^-,c}^{-}$), heterotrophic denitrification with nitrate ($NO_{3\ HD,NO_3^-,c}^{-}$), methanol oxidation ($NO_{3\ MO,c}^{-}$), and heterotrophic partial denitrification with nitrate reduction to nitric oxide by *Lewinellaceae* ($NO_{3\ Lew,c}^{-}$) as shown in Eq.3.

$$NH_{4\ eff}^{+} = NH_{4\ inf}^{+} + NH_{4\ O.O,p}^{+} + NH_{4\ Fer.,p}^{+} + NH_{4\ HD,NO_2^-,p}^{+} + NH_{4\ HPD,NO_3^-,p}^{+} + NH_{4\ DNRA,p}^{+} + NH_{4\ SR,p}^{+} + NH_{4\ HD,NO_3^-,p}^{+} - NH_{4\ AAO,c}^{+} - NH_{4\ N.O,c}^{+} - NH_{4\ APD,c}^{+} - NH_{4\ Anx,c}^{+} - NH_{4\ AD,NO_3^-,c}^{+} - NH_{4\ MO,c}^{+} \quad \text{(Eq.1)}$$

$$NO_{2\ eff}^{-} = NO_{2\ inf}^{-} + NO_{2\ AAO,p}^{-} + NO_{2\ HPD,NO_3^-,p}^{-} - NO_{2\ HD,NO_2^-,c}^{-} - NO_{2\ N.O.,c}^{-} - NO_{2\ Anx.,c}^{-} \quad \text{(Eq.2)}$$

$$NO_{3\ eff}^{-} = NO_{3\ inf}^{-} + NO_{3\ N.O.,p}^{-} + NO_{3\ Anx.,p}^{-} - NO_{3\ APD.,c}^{-} - NO_{3\ AD.,NO_3^-,c}^{-} - NO_{3\ DNRA,c}^{-} - NO_{3\ HPD,NO_3^-,c}^{-} - NO_{3\ HD,NO_3^-,c}^{-} - NO_{3\ MO,c}^{-} - NO_{3\ Lew,c}^{-} \quad \text{(Eq.3)}$$

Operational parameters and influent nitrogen compounds concentrations were the input process variables in data sets of Group 1, while data sets in Group 2 also included microbial community abundance percentages as input process variables to determine the influence of microbiome on the prediction of effluent nitrogen compounds concentrations. The statistical features of data sets in group 1 and 2 are listed in **Table 1**.

Final data sets in Group 1 to predict effluent nitrogen compounds concentrations resulted in a total of 1,660 data points, while final data sets in group 2 consisted of 10,956 data points. Same-day information was classified as "measured" or "calculated" depending on whether data reflect actual data (40 days) or theoretically calculated data (126 days). Both Group 1 and Group 2 have 166 groups of data.

2.3 Model development and ranking

Random forest (RF) and Support Vector Machine (SVM) are employed for training and testing of data sets in groups 1 and 2. These algorithms are selected mainly due to their capability of capturing non-linear relationships in highly complex datasets. RF is an ensemble method which consists of a user-specified number of decision trees, usually Classification and Regression Trees, each of which is trained on a sub-sample of the data.[58,59] Decision trees such as CART are formed by splitting features to make decision so that each split maximizes the reduction of variance in regression tasks and Gini impurity in classification tasks. SVM models for regression tasks, also known as Support Vector Regression (SVR) models, are trained by minimizing a convex cost function that resembles the discrepancy between the actual and predicted values.[60,61] For comparison purposes, linear regression serves as a baseline to evaluate the regression performance of these more complex models.

The algorithms presented in this paper are developed, trained, and tested using Python 3.6, and the scikit-learn,[62] TensorFlow™,[63] and KERAS[64] packages. The hyperparameters of the studied machine learning algorithms were determined using guidelines provided in Asghari et al.[65] and Wang et al.[66] Optimal values for hyperparameters corresponding to machine learning algorithms are shown in **Table S10**. All models were trained with 126 theoretically calculated same-day data and tested with random selections of 20 measured same-day data. This approach was designed to assess whether models trained on calculated data could achieve comparable

performance to those trained on real data. Prior to training neural networks models, all features were scaled using MinMax function, so they range between 0 and 1. Correlation coefficient (R), mean absolute error (MAE), and mean squared error (MSE) were used to rank the performance of the models developed by the machine learning methods.

2.4 Data set size and feature selection analysis

Data size was analyzed in depth and breadth. In depth, because the machine learning algorithm with the best performance for the data sets was selected to assess whether calculating missing values through theoretical calculations aids in the performance of the models by varying the training data set size from 0% (0 same-day data) to 100% (146 same-day data). The R of the predictions were benchmarked against traditional linear regression. In breadth, because microbial community abundance percentages were included as input process variables for the analysis of data sets in group 2. Microbial analysis is labor intensive and time-consuming; thus, it is vital to determine whether obtaining microbiome data improves predictions. In addition, the identification of the relative importance (RI) of input process parameters in the prediction of effluent nitrogen compounds concentrations was performed by the machine learning algorithm with the best performance. Data sets of Group 1 have nine input process parameters available, while data sets in Group 2 have over sixty input process parameters. To obtain a better generalized model, the most relevant parameters were identified. Data sets and machine learning codes are available online.[1]

[1] https://github.com/vd1371/ML-PNAFMBR

3. RESULTS AND DISCUSSION

3.1 Prediction performance

Predicted values for effluent ammonium, nitrite, and nitrate concentrations were compared with the training, cross validation, and test sets. To simulate the effluent nitrogen compounds concentrations in the PN/AFMBR, the optimal values for the hyperparameters corresponding to machine learning algorithms were identified.

Table 2. Performance of machine learning algorithms on testing data set including microbial community abundance percentages.

Output process variable	Criteria	LR	RF	SVM
Ammonium	R	0.008	0.894	0.794
	MAE	5.65	2.022	2.540
	MSE	31.96	7.671	13.793
Nitrite	R	0.189	0.722	0.351
	MAE	4.18	0.891	1.389
	MSE	25.84	1.648	5.069
Nitrate	R	0.247	0.802	0.808
	MAE	5.15	1.787	1.857
	MSE	56.14	6.060	6.667
Overall Rank		III	I	II

Performance evaluations in the test data sets aid to determine the robustness of the predictive models. **Table 2** displays the R, MAE and MSE values obtained by the three machine learning methods estimating the difference between effluent and influent ammonium, nitrite and nitrate concentrations. Ranking of each algorithm based on three evaluation criteria is also provided in **Table 2**. Based on the results, change in ammonium concentrations are satisfactorily fitted by random forest with close values for R (0.89), MAE (2.02) and MSE (7.67). For nitrite concentrations, random forest outperforms the other algorithms by providing the highest R (0.72), and the lowest MAE (0.89) and MSE (1.65) errors. For the difference in nitrate concentrations, SVM, presented the highest R of 0.81, while RF presented the lowest MAE of 1.79, and the lowest MSE of 6.06, respectively. Comparison of prediction performance with

311 theoretical-empirical models is not conducted because of the lack of studies modeling change
312 in nitrogen compounds concentrations in the integrated PN/A FMBR with GAC fluidization.
313 Indeed, there are some studies that focus on PN/A process without membranes in their
314 configurations,[17,34,35,67] or MBRs for activated sludge.[31,68] However, some of these studies
315 focus on mathematical predictions without real data,[17,34] or they lack model validation on test
316 datasets. [31,35,67,68]

317 Overall, random forest presented the best performance across all nitrogen compounds verifying
318 its robustness to predict the profile of effluent nitrogen compounds concentrations in the PN/A
319 FMBR compared to other algorithms. The change in nitrogen compounds concentrations was
320 accurately predicted within one standard deviation of the data sets in group 1. Unlike SVM
321 which is trained on the whole training set, elements (trees) of random forest can overlook
322 mistaken data entries, noises, or anomalies. Moreover, data scaling is not required for the pre-
323 processing of random forest, unlike SVM. Owing to these facts, training random forest is easier
324 and more intuitive for engineers. With this flexibility provided by random forest, or ensemble
325 methods in general, practitioners could, theoretically, train more accurate models on smaller
326 datasets.

327 **3.2 Data size analysis**

328 The comparison of correlation coefficients (R) between predicted concentrations using
329 different quantities of training data for random forest (RF) and linear regression (LR) is
330 illustrated in **Figure 2**. R^2 at different sizes of training datasets is included in Supporting
331 Information, Figure S2. The dataset consists of 166 days in total, including 40 days of measured
332 same-day data and 126 days of theoretically calculated data for ammonium, nitrite, and nitrate.
333 Test sets comprised 20 measured same-day data, while training sets varied from 0 to 146 days,

with the remaining 20 same-day data in the training set representing approximately 14% of the data (shown by the blue vertical line).

For datasets without microbial community information (**Figure 2A**), the correlation coefficients stabilized from 50% of the data in the training set onwards. RF consistently outperformed LR, with correlation ranges of 0.803 to 0.859 for ammonium, 0.665 to 0.772 for nitrite, and 0.897 to 0.922 for nitrate concentrations, while LR's R values ranged from 0.663 to 0.726 for ammonium, 0.664 to 0.701 for nitrite, and 0.527 to 0.677 for nitrate. The highest R for ammonium (0.859) and nitrate (0.922) were achieved at 50% of the training data using RF, while the highest R for nitrite (0.772) was obtained at 80%. Notably, at 40% of the training set, the R value for LR (0.819) slightly exceeded that of RF (0.815) for ammonium, suggesting that better tuning of RF hyperparameters could further improve its performance. For nitrite, LR outperformed RF between 15% and 55% of the training set, while RF showed more stable performance after 60% of the training data.

The performance for nitrate predictions was more consistent with RF, maintaining correlation values between 0.860 and 0.922, while LR exhibited decreasing R values with increasing data size, dropping from 0.527 to 0.882. The lowest R values for both algorithms occurred when the training set was limited to the 20 measured same-day data. However, training sets containing 50% to 60% of the data (73 to 88 same-day data) reached acceptable R values using RF, highlighting the importance of calculated values for missing data.

When microbial community information was included (**Figure 2B**), the model performance improved significantly. For ammonium, the R value increased from 0.894 to 0.938, mean absolute error (MAE) decreased from 2.022 to 1.572, and mean squared error (MSE) dropped from 7.671 to 4.508, as presented in **Table S11**. For nitrite, R improved from 0.722 to 0.851, with MAE decreasing from 0.891 to 0.651, and MSE from 1.648 to 1.346. Similarly, nitrate

predictions improved, with R increasing from 0.802 to 0.942, MAE decreasing from 1.787 to 0.917, and MSE from 6.060 to 1.915. This demonstrates that including microbial community data enhances model accuracy by providing more comprehensive system information.

With the additional microbial data, RF continued to show strong performance, with consistent R values ranging from 0.781 to 0.959 for ammonium, 0.621 to 0.846 for nitrite, and 0.920 to 0.964 for nitrate, compared to LR. The highest R values were obtained at 90%, 100%, and 50% of the data for ammonium, nitrite, and nitrate, respectively. For ammonium and nitrate, R values greater than 0.90 were achieved with only 50% and 10% of the training data, respectively, while nitrite predictions required larger data sets to reach an R value above 0.80.

The hybrid theoretical-machine learning approach leverages stoichiometry and kinetic models to fill in missing data for empirical analysis, offering better predictions for nitrogen concentrations. Besides, the comparison with LR indicates that ML-Theoretical Model deliver a better performance even under constraints of limited data, and the predictive performance improves faster than LR when the data size increases. While traditional theoretical models may overlook complex microbial interactions, integrating microbial community data allowed this approach to identify empirical patterns and interactions that theoretical calculations alone might miss. This combination of methods leads to a more comprehensive understanding of system dynamics, as demonstrated by the improved predictions over the First Principal Model. Detailed comparisons between the ML-theoretical approach and the First Principal Model, including the effects on prediction outcomes, are provided in **Supporting Method S3 and Tables S2-S7**.

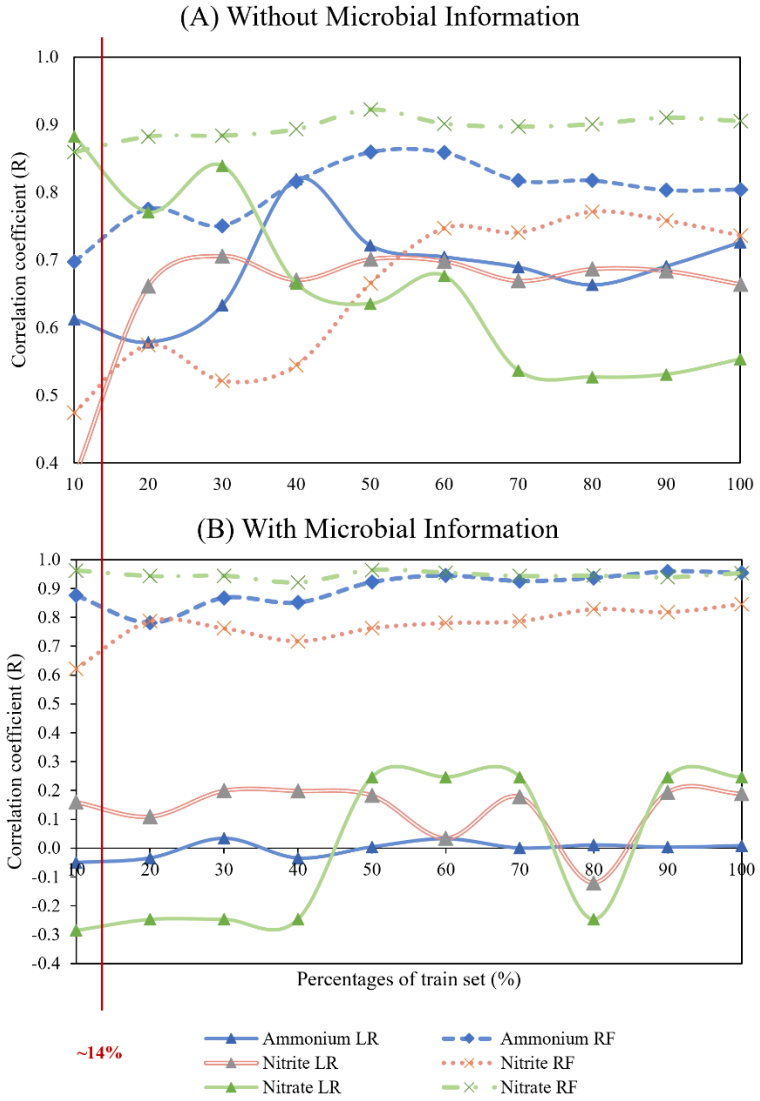


Figure 2. Correlation coefficients at different sizes of training data sets as predicted by linear regression (LR) and random forest (RF). Red vertical line represents the percentage of test data sets (~14%). (A) datasets without microbial community information; (B) datasets include microbial community information.

3.3 Identification of Key Environmental Factors

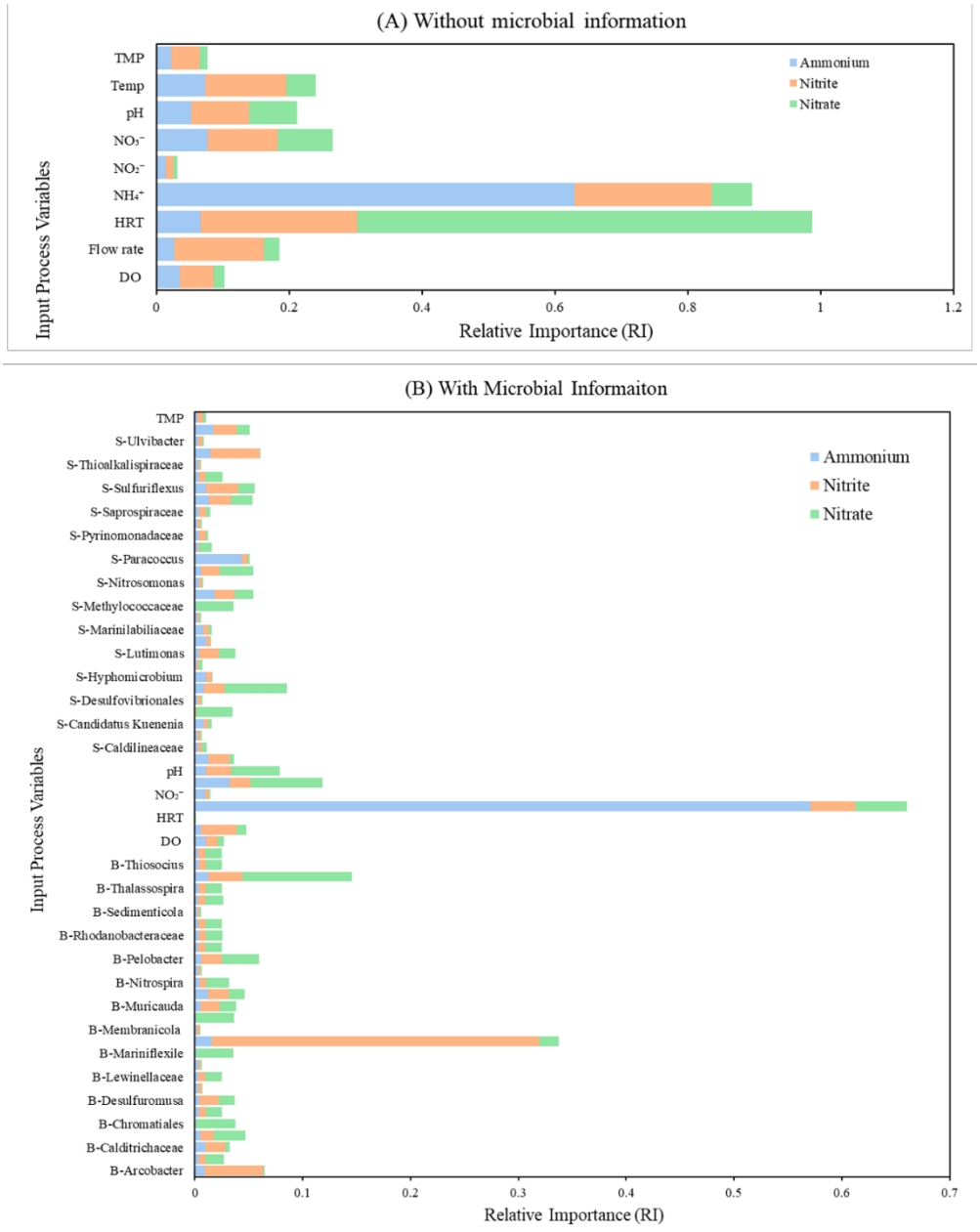
Variations from influent to effluent nitrogen compounds concentration depends on several parameters. For the change in ammonium concentrations in Group 1 (**Figure 3A**), pollution load (ammonium with RI of 0.630 and nitrate concentration in the influent with RI of 0.077), and temperature (RI = 0.074) were the dominant parameters influencing the prediction values.

These three parameters were also influential when considering microbial community abundances (Group 2) with RI of 0.570 for ammonium concentration in the influent, of 0.032 for nitrate concentration in the influent, and of 0.017 for temperature.

Temperature, known to be critical for nitrification, is consistent with previous findings [40], especially in relation to the optimal growth rate of *Nitrosomonas* (25-35ºC). Feature selection from random forest algorithms also revealed that nitrate concentration in the influent significantly impacts ammonium concentration predictions. Feature selection from random forest algorithms also revealed that nitrate concentration in the influent significantly impacts ammonium concentration predictions.

For nitrite concentrations, HRT (RI = 0.236), influent ammonium (RI = 0.207), flow rate (RI = 0.135), temperature (RI = 0.122), and influent nitrate (RI = 0.106) were identified as key factors in group 1. These environmental factors were less dominant when microbial data was incorporated as shown in group 2 where influent ammonium concentration (RI = 0.042), flow rate (RI = 0.032), and temperature (RI = 0.021) remained as influential parameters (**Figure 3B**). The relative importance of HRT decreased because microbes – especially in biofilm – depend in solids retention time rather than HRT

For nitrate concentrations, HRT (RI = 0.685), nitrate concentration in the influent (RI = 0.083), and pH (RI = 0.072) emerged as the most critical parameters in Group 1. In group 2, nitrate concentration in the influent (RI = 0.067), ammonium concentration in the influent (RI = 0.047), and pH (RI = 0.045) are the key operational parameters identified (**Figure 3B**). Similar to nitrite data sets, the relative importance of HRT decreased because microbes are not controlled by HRT but solids retention time. Microbiomes removing nitrate presented higher influence in nitrate concentrations than microbiomes producing nitrate.



412

413 **Figure 3**. Relative Importance of influential parameters for effluent concentrations: Ammonium, Nitrite
414 concentrations, Nitrate (A) without microbial information (B) with microbial information.

3.4 Discussion of Microbial Interactions

In Group 2 **(Figure 3B)**, which considered microbial community data, specific bacteria played critical roles in nitrogen transformations. For ammonium concentrations, *Paracoccus* (RI = 0.043), *Muricauda* (RI = 0.018), *Thiosocius* (RI = 0.015), *Sedimenticola* (RI = 0.013), *Arcobacter* (RI = 0.012), and *Hyphomicrobium* (RI = 0.011) in suspension sludge, and *Marinilabiliaceae* (RI = 0.016), *Thioalkalispiraceae* (RI = 0.013) and *Nitrosomonas* (RI = 0.013) in the biofilm were also identified as influential factors in the change of ammonium concentrations in group 2. Given that ammonium is removed during the first step of nitrification as performed by ammonium oxidizing bacteria (i.e., *Nitrosomonas*), influent ammonium concentration and temperature are expected as key influential parameters for change in ammonium concentrations predictions. *Hyphomicrobium* [70], *Sedimenticola* [71], *Thiosocius* [71], and *Arcobacter* [72] in suspension sludge; and *Marinilabiliaceae* [73] and *Thioalkalispiraceae*[74] in the biofilm of GAC reduce nitrate while removing or producing low quantities of ammonium. Positive correlations were previously found through metagenomic sequencing analyses between *Hyphomicrobium* and *Paracoccus*, and *Sedimenticola* and *Muricauda*.[49] *Paracoccus*[75] and *Muricauda* remove organics while producing low quantities of ammonium concentrations.

Nitrosomonas (ammonium oxidizing bacteria) with nitrite production, and *Nitrospira* (NOB) and *Candidatus* Kuenenia (Anammox bacteria) with nitrite removal were expected to lead as key microbiomes in the change of nitrite concentrations. However, random forest algorithm suggests that partial denitrifiers have higher influence than *Nitrosomonas* (RI = 0.019 in biofilm), *Nitrospira* (RI = 0.017 in suspension sludge) and *Anammox* bacteria (RI = 0.012 in biofilm) in this PN/AFMBR. This could be due to several reasons. First, partial denitrifiers may be more adaptable to the varying environmental conditions within the reactor, allowing them to maintain higher activity levels [76]. Second, the metabolic flexibility of partial denitrifiers

might enable them to utilize a broader range of substrates including both inorganic and organic carbon, thus exerting a more significant influence on the overall nitrogen removal process. Lastly, interactions between partial denitrifiers and other microbial communities could enhance their relative importance, possibly through synergistic relationships that boost their effectiveness in nitrogen compound transformation. Further investigation is needed to elucidate these interactions and their impact on the system's performance.

For nitrite concentration, *Marinilabiliaceae* (RI = 0.304) and *Arcobacter* (RI = 0.055) in the biofilm, and *Thiosocius* (RI = 0.045) in the suspensions sludge were the dominant factors. *Marinilabiliaceae* [73] and *Arcobacter* [72] have been identified as heterotrophic partial denitrifiers with nitrate reduction to nitrite. Positive correlations between *Thiosocius* and *Arcobacter* in suspension sludge, and with *Marinilabiliaceae* through intermediate species (i.e., *Rhodanobacteraceae*) in biofilm were identified through previous metagenomic sequencing analysis.[49] *Sulfuriflexus*[77] and *Sedimenticola*[71] have similar physiological functions to *Thiosocius*[78] and have been identified as autotrophic partial denitrifiers with nitrate reduction. *Sulfuriflexus* (RI = 0.030) and *Sedimenticola* (RI = 0.020) in the suspension sludge were also key influential microbiome. Nitrite has not been found to be consumed nor produced by *Thioalkalispiraceae*[74] (RI = 0.031) and *Pelobacter*[77] (RI = 0.019); yet random forest method found these species in the biofilm to be influential in the change of nitrite concentrations.

In nitrate predictions, *Thioalkalispiraceae* (RI = 0.102) and *Chromatiales* (RI = 0.037) in the biofilm, and *Chromatiales* (RI = 0.034) in the suspension sludge are autotrophic denitrifiers removing nitrate.[74,79] *Chromatiales* demonstrated positive correlation to *Methylococcaceae* as investigated by metagenomics sequencing analysis.[80] *Methylococcaceae* present in biofilm (RI = 0.035) and suspension sludge (RI = 0.035) reduce nitrate to nitric oxide through methanol oxidation.[81,82] *Desulfuromusa* (RI = 0.058) in suspension sludge, and *Pelobacter* (RI = 0.034) in biofilm are key microbiome consuming nitrate through heterotrophic dissimilatory nitrate

reduction to ammonium.[77,83] *Nitrospira* in suspension sludge (RI = 0.031), *Candidatus* Kuenenia (RI = 0.029) and *Nitrospira* in biofilm (RI = 0.021) are nitrate producing bacteria which influenced the change in nitrate concentration as expected.[5] This also aligns with the understanding that maintaining stable suppression of nitrite-oxidizing bacteria (NOB) is crucial for achieving optimal PN/A performance.[84] *Mariniflexile*[85] in biofilm (RI = 0.035) does not consume nor produce nitrate but presented positive correlation to *Thioalkalispiraceae* in biofilm which might indicate synergistic relationships among them.

Previous PN/AFMBR model groups microbiomes depending on their physiological functions.[49] For example, heterotrophic partial denitrifiers with nitrate reduction to nitrite within the PN/AFMBR included *Arcobacter*, *Ulvibacter*, *Lutimonas*, and *Marinilabiliaceae*. This hybrid approach including machine learning algorithms provided in-depth analysis by identifying that among all heterotrophic partial denitrifiers, *Marinilabiliaceae* and *Arcobacter* in the biofilm are the key influential microbiomes affecting the change of nitrite concentrations in saline sewage. Microbiomes identified at the genus level obtained by 16S rRNA gene sequencing are not handled by traditional biological models. Yet, data sets for statistical-empirical approach would contain missing data without fixed theoretical calculations as provided in traditional models. Thus, the integration of both approaches overcomes each other's shortcomings and identifies hidden microbial interactions.

3.5 Environmental Implications

This study proposed a hybrid theoretical-machine learning approach framework, highlighting the feasibility of leveraging limited microbial data for accurate predictions of nitrogen removal performance. By integrating microbial community data, the framework provides actionable insights into the key environmental parameters and microbial interactions that influence nitrogen removal efficiency. Moreover, this research underscores the importance of improving

488 microbial data collection methods and integrating them into predictive models. Enhanced
489 microbial monitoring in full-scale systems could facilitate more precise control over nitrogen
490 removal processes, leading to better environmental outcomes. By bridging the gap between
491 pilot-scale experimentation and real-world applications, this study paves the way for the
492 development of advanced decision-making tools tailored to biological nitrogen removal
493 systems.

494 **3.6 Limitations & future works**

495 This hybrid method of theoretical calculations and machine learning algorithms has certain
496 limitations. First, theoretical models depend on known microbiomes and steady-state
497 assumptions, which can miss dynamic changes in microbial populations. Despite advances in
498 meta-omics, about 27-60% of the microbiomes in the PN/A FMBR samples were "unclassified."
499 As new bacteria are discovered, they can be incorporated into models like the TEE, improving
500 predictions by updating missing data. Steady-state models also struggle with daily fluctuations
501 in influent characteristics, requiring random values to fill gaps for unmeasured periods. Besides,
502 machine learning algorithms need large datasets to train effectively. In this study, 73-88 same-
503 day measurements were necessary to achieve high correlation predictions for nitrogen
504 compounds, but collecting and analyzing microbial DNA from such large samples can be
505 expensive and time-consuming. However, the hybrid approach mitigates individual limitations
506 by leveraging the strengths of both methods, resulting in acceptable prediction errors using the
507 random forest algorithm.

508 Given that random forest algorithms presented well fitted predictions of changes in nitrogen
509 compounds concentrations of a PN/A FMBR, this hybrid theoretical-machine learning
510 approach could be tested for other biological wastewater and sludge treatment processes.
511 Furthermore, hidden microbiomes relationships could be further investigated. Future research

should explore hidden microbial relationships, particularly interactions between key species like *Thiosocius* and *Marinilabiliaceae* in the biofilm, and further investigate how autotrophic denitrifiers like *Thioalkalispiraceae* interact with heterotrophic reducers like *Pelobacter*. Additionally, the study identified 11 key genera in both suspension sludge and biofilm that influence nitrogen concentrations, suggesting the need for further exploration of their interactions with operational parameters. Future research should focus on validating this methodology with additional datasets from full-scale systems to ensure its robustness and generalizability. This validation will help refine the model's predictive accuracy and further optimize operational strategies. As full-scale systems often encounter similar challenges, such as data gaps and fluctuating influent characteristics, this study provides a promising approach to addressing these limitations.

4. CONCLUSIONS

Traditional sewage treatment processes often face difficulties due to incomplete or invalid data, which hampers effective optimization and scaling. By utilizing a hybrid theoretical-machine learning approach, this study successfully bridges these gaps by calculating missing values and revealing hidden microbial interactions that traditional methods overlook. The main conclusions are as follows:

1) Random forest consistently outperformed other algorithms, showing robust prediction accuracy for ammonium, nitrite, and nitrate concentrations, with correlation coefficients of 0.89, 0.72, and 0.80, respectively.
2) Including microbial community abundance data significantly improved model performance, revealing critical microbial interactions, particularly for nitrogen transformation and partial denitrification.

3) Key parameters such as temperature, pH, and microbial communities (e.g., heterotrophic denitrifiers like *Paracoccus*, *Marinilabiliaceae* and autotrophic denitrifiers, such as *Thioalkalispiraceae*) were identified as essential for efficient PN/AFMBR performance.

5. ACKNOWLEDGEMENTS

The authors thank the Hong Kong Research Grants Council- University Grants Committee (Grant No. 15252916 and UGC/GEN/456/08), and the Research Institute for Sustainable Urban Development (RISUD) for their financial support. Declarations of interest: none.

6. SUPPORTING INFORMATION

The "Supporting Information" file contains the assumptions for missing values of nitrogen compounds concentrations in the effluent, explanation of dates and frequency of microbial community analysis sampling, prediction of effluent parameters based on mathematical models, statistical description of microbial community abundance percentages data, optimal values of hyperparameters for machine learning algorithms and performance of machine learning algorithms on test set including microbial community abundance percentages.

7. REFERENCES

(1) McCarty, P. L. What is the Best Biological Process for Nitrogen Removal: When and Why? Environ. *Sci Technol* **2018**, *52*, 3835–3841.

(2) Chen, R.; Ji, J.; Chen, Y.; Takemura, Y.; Liu, Y.; Kubota, K.; Ma, H.; Li, Y.-Y. Successful Operation Performance and Syntrophic Micro-Granule in Partial Nitritation and Anammox Reactor Treating Low-Strength Ammonia Wastewater. *Water Res.* **2019**, *155*, 288–299. https://doi.org/10.1016/j.watres.2019.02.041.

(3) Huang, X.; Sun, K.; Wei, Q.; Urata, K.; Yamashita, Y.; Hong, N.; Hama, T.; Kawagoshi, Y. One-Stage Partial Nitritation and Anammox in Membrane Bioreactor. *Env. Sci Pollut Res* **2016**, *23*, 11149–11162.

(4) Huang, X.; Mi, W.; Chan, Y. H.; Singh, S.; Zhuang, H.; Leu, S. Y.; Li, X.; Li, X.; Lee, P. H. C-N-S Synergy in a Pilot-Scale Mainstram Anammox Fluidized-Bed Membrane

Bioreactor for Treating Chemically Enhanced Primary Treatment Saline Sewage. *Water Res.* **2023**, *229* (1), 119475.

(5) Jetten, M. S. M.; Horn, S. J.; Loosdrecht, M. C. M. Towards a More Sustainable Municipal Wastewater Treatment System. *Water Sci Technol* **1997**, *35*, 171–180.

(6) Mulder, A.; Graaf, A. A.; Robertson, L. A.; Kuenen, J. G. Anaerobic Ammonium Oxidation Discovered in Nitrifying Fluidized Bed Reactor. *Microbiol Ecol* **1995**, *16*, 177–184.

(7) Strous, M.; Heijnen, J. J.; Kuenen, J. G.; Jetten, M. S. M. The Sequencing Batch Reactor as a Powerful Tool for the Study of Slowly Growing Anaerobic Ammonium-Oxidizing Microorganisms. *Appl. Microbiol. Biotechnol.* **1998**, *50* (5), 589–596. https://doi.org/10.1007/s002530051340.

(8) Lee, P.; Kwak, W.; Bae, J.; Mccarty, P. L. The Effect of SRT on Nitrate Formation during Autotrophic Nitrogen Removal of Anaerobically Treated Wastewater. *Water Sci Technol* **2013**, *68*, 1751–1756.

(9) Abdelmelek, S. B.; Greaves, J.; Ishida, K. P.; Cooper, W. J.; Song, W. Removal of Pharmaceutical and Personal Care Products from Reverse Osmosis Retentate Using Advanced Oxidation Processes. *Environ. Sci. Technol.* **2011**, *45* (8), 3665–3671. https://doi.org/10.1021/es104287n.

(10) Katsou, E.; Malamis, S.; Loizidou, M. Performance of a Membrane Bioreactor Used for the Treatment of Wastewater Contaminated with Heavy Metals. *Bioresour Technol* **2011**, *102*, 4325–4332.

(11) Wu, B.; Li, Y.; Lim, W.; Lin, S.; Guo, Q.; Fane, A. G. Single-Stage versus Two-Stage Anaerobic Fluidized Bed Bioreactors in Treating Municipal Wastewater : Performance, Foulant Characteristics , and Microbial Community. *Chemosphere* **2017**, *171*, 158–167.

(12) Chen, F.; Qian, Y.; Cheng, H.; Shen, J.; Qin, Y.; Li, Y.-Y. Recent Developments in Anammox-Based Membrane Bioreactors: A Review. *Sci. Total Environ.* **2023**, *857*, 159539. https://doi.org/10.1016/j.scitotenv.2022.159539.

(13) Li, J.; Li, J.; Peng, Y.; Wang, S.; Zhang, L.; Yang, S.; Li, S. Insight into the Impacts of Organics on Anammox and Their Potential Linking to System Performance of Sewage Partial Nitrification-Anammox (PN/A): A Critical Review. *Bioresour. Technol.* **2020**, *300*, 122655. https://doi.org/10.1016/j.biortech.2019.122655.

(14) Wang, Z.; Zheng, M.; Duan, H.; Yuan, Z.; Hu, S. A 20-Year Journey of Partial Nitritation and Anammox (PN/A): From Sidestream toward Mainstream. *Environ. Sci. Technol.* **2022**, *56* (12), 7522–7531. https://doi.org/10.1021/acs.est.1c06107.

(15) Zhang, X.; Zhang, X.; Chen, J.; Wu, P.; Yang, Z.; Zhou, L.; Zhu, Z.; Wu, Z.; Zhang, K.; Wang, Y.; Ruth, G. A Critical Review of Improving Mainstream Anammox Systems: Based on Macroscopic Process Regulation and Microscopic Enhancement Mechanisms. *Environ. Res.* **2023**, *236*, 116770. https://doi.org/10.1016/j.envres.2023.116770.

(16) Harrou, F.; Dairi, A.; Sun, Y.; Senouci, M. Statistical Monitoring of a Wastewater Treatment Plant: A Case Study. *J Env. Manage* **2018**, *223*, 807–814.

(17) Hubaux, N.; Wells, G.; Morgenroth, E. Impact of Coexistence of Flocs and Biofilm on Performance of Combined Nitritation-Anammox Granular Sludge Reactors. *Water Res* **2015**, *68*, 127–139.

(18) Rodriguez-Sanchez, A.; Muñoz-Palazon, B.; Hurtado-Martinez, M.; Maza-Marquez, P.; Gonzalez-Lopez, J.; Vahala, R.; Gonzalez-Martinez, A. Microbial Ecology Dynamics of a Partial Nitritation Bioreactor with Polar Arctic Circle Activated Sludge Operating at Low Temperature. *Chemosphere* **2019**, *225*, 73–82.

(19) Al-Hazmi, H. E.; Lu, X.; Grubba, D.; Majtacz, J.; Badawi, M.; Mąkinia, J. Sustainable Nitrogen Removal in Anammox-Mediated Systems: Microbial Metabolic Pathways, Operational Conditions and Mathematical Modelling. *Sci. Total Environ.* **2023**, *868*, 161633. https://doi.org/10.1016/j.scitotenv.2023.161633.

(20) Sharifshourjeh, M.; Kowal, P.; Xi, L.; Xie, L.; Drewnowski, J. Development of Strategies for AOB and NOB Competition Supported by Mathematical Modeling in Terms of Successful Deammonification Implementation for Energy-Efficient WWTPs. *Processes* **2021**, *9*, 562. https://doi.org/10.3390/pr9030562.

(21) Kosgey, K.; Zungu, P. V.; Kumari, S.; Bux, F. Critical Review of Process Control Strategies in Anammox-Mediated Nitrogen Removal Systems. *J. Environ. Chem. Eng.* **2022**, *10* (4), 108068. https://doi.org/10.1016/j.jece.2022.108068.

(47) Lamrini, B.; Lakhal, E.-K.; Le Lann, M.-V.; Wehenkel, L. Data Validation and Missing Data Reconstruction Using Self-Organizing Map for Water Treatment. *Neural Comput. Appl.* **2011**, *20* (4), 575–588. https://doi.org/10.1007/s00521-011-0526-5.

(23) Vannecke, T. P. W.; Bernet, N.; Winkler, M. K. H.; Santa-Catalina, G.; Steyer, J. P.; Volcke, E. I. P. Influence of Process Dynamics on the Microbial Diversity in a Nitrifying Biofilm Reactor: Correlation Analysis and Simulation Study. *Biotechnol Bioeng* **2016**, *113*, 1962–1974.

(24) Gujer, W.; Henze, M.; Mino, T.; Loosdrecht, M. Activated sludge model no.3. *Water Sci Technol* **1999**, *39*, 183–193.

(25) Henze, M.; Grady Jr, L.; Gujer, W.; Marais, G.; Matsuo, T. Activated Sludge Model No 1. *Wat Sci Technol* **1987**, *29*.

(26) Batstone, D. J.; Keller, J.; Angelidaki, I.; Kalyuzhnyi, S. V.; Pavlostathis, S. G. *Ananerobic digestion model no.1 (ADM1)*; Water Science and technology, 2002; Vol. 45(10).

(27) Reichert, P. AQUASIM 2.0 - User Manual. In *Computer Program for the Identification and Simulation of Aquatic Systems*; 1998.

(28) Water and Wastewater Treatment Modeling and Simulation Software| Hydromantis VERSION 8.0, 2019. https://www.hydromantis.com/ (accessed 2023-12-28).

(29) *EnviroSim – Wastewater Modeling Software*. https://envirosim.com/ (accessed 2023-12-28).

(30) Cho, J. W.; Ahn, K. H.; Lee, Y. H.; Lim, B. R.; Kim, J. Y. Investigation of Biological and Fouling Characteristics of Submerged Membrane Bioreactor Process for Wastewater Treatment by Model Sensitivity Analysis. *Water Sci Technol* **2004**, *49*, 245–254.

(31) Sarioglu, M.; Insel, G.; Artan, N.; Orhon, D. Stoichiometric and Kinetic Evaluation of Simultaneous Nitrification and Denitrification in a Membrane Bioreactor at Steady State. *J Chem Technol Biotechnol* **2011**, *86*, 798–811.

(32) Mehrdad, M.; Park, H.; Chandran, K.; Ramalingam, K.; Fillos, J. Biofilm Population Diversity and Distribution in an Anammox MBBR Pilot: Molecular Analysis and

Mathematical Modeling; Water Environment Federation, 2016; Vol. 2016/11. https://doi.org/10.2175/193864716819706707.

(33) Pérez, J.; Lotti, T.; Kleerebezem, R.; Picioreanu, C.; Loosdrecht, M. C. M. Outcompeting Nitrite-Oxidizing Bacteria in Single-Stage Nitrogen Removal in Sewage Treatment Plants: A Model-Based Study. *Water Res* **2014**, *66*, 208–218.

(34) Rittmann, B. E.; Boltz, J. P.; Brockmann, D.; Daigger, G. T.; Morgenroth, E.; Sørensen, K. H.; Takács, I.; Loosdrecht, M.; Vanrolleghem, P. A. A Framework for Good Biofilm Reactor Modeling Practice (GBRMP). Water Sci. *Technol* **2018**, *77*, 1149–1164.

(35) Volcke, E. I. P.; Picioreanu, C.; Baets, B.; Loosdrecht, M. C. M. The Granule Size Distribution in an Anammox-Based Granular Sludge Reactor Affects the Conversion-Implications for Modeling. *Biotechnol Bioeng* **2012**, *109*, 1629–1636.

(36) Kim, M.; Kim, Y.; Kim, H.; Piao, W.; Kim, C. Evaluation of the K-Nearest Neighbor Method for Forecasting the Influent Characteristics of Wastewater Treatment Plant. *Front Env. Sci Eng* **2016**, *10*, 299–310.

(37) Newhart, K. B.; Holloway, R. W.; Hering, A. S.; Cath, T. Y. Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Res.* **2019**, *157*, 498–513. https://doi.org/10.1016/j.watres.2019.03.030.

(38) Park, M.; Anumol, T.; Snyder, S. A. Modeling Approaches to Predict Removal of Trace Organic Compounds by Ozone Oxidation in Potable Reuse Applications. *Env. Sci Water Res Technol* **2015**, *1*, 699–708.

(39) Qiao, J.; Hu, Z.; Li, W. Soft Measurement Modeling Based on Chaos Theory for Biochemical Oxygen Demand. *Water* **2016**, *8(12)* (581).

(40) Verma, A.; Wei, X.; Kusiak, A. Predicting the Total Suspended Solids in Wastewater: A Data-Mining Approach. *Eng Appl Artif Intell* **2013**, *26*, 1366–1372.

(41) Vega, P. T.; Jaramillo-Moran, M. A. Obtaining Key Parameters and Working Conditions of Wastewater Biological Nutrient Removal by Means of Artificial Intelligence Tools. *Water* **2018**, *10*.

(42) Awolusi, O. O.; Nasr, M.; Kumari, S.; Bux, F. Artificial Intelligence for the Evaluation of Operational Parameters Influencing Nitrification and Nitrifiers in an Activated Sludge Process. *Microb Ecol* **2016**, *72*, 49–63.

(43) Zhao, L.; Wu, Q. Review and Expectation of Artificial Intelligent System for Wastewater Treatment. *Appl Mech Mater* **2013**, *422*, 237–241.

(44) Wang, D.; Thunéll, S.; Lindberg, U.; Jiang, L.; Trygg, J.; Tysklind, M. Towards Better Process Management in Wastewater Treatment Plants: Process Analytics Based on SHAP Values for Tree-Based Machine Learning Methods. *J. Environ. Manage.* **2022**, *301*, 113941. https://doi.org/10.1016/j.jenvman.2021.113941.

(45) Xu, Y.; Zeng, X.; Bernard, S.; He, Z. Data-Driven Prediction of Neutralizer pH and Valve Position towards Precise Control of Chemical Dosage in a Wastewater Treatment Plant. *J. Clean. Prod.* **2022**, *348*, 131360. https://doi.org/10.1016/j.jclepro.2022.131360.

(46) Olsson, G. Instrumentation, Control and Automation in the Water Industry – State-of-the-Art and New Challenges. *Water Sci. Technol.* **2006**, *53* (4–5), 1–16. https://doi.org/10.2166/wst.2006.097.

(47) Lamrini, B.; Lakhal, E.-K.; Le Lann, M.-V.; Wehenkel, L. Data Validation and Missing Data Reconstruction Using Self-Organizing Map for Water Treatment. *Neural Comput. Appl.* **2011**, *20* (4), 575–588. https://doi.org/10.1007/s00521-011-0526-5.

(48) Duarte, M. S.; Martins, G.; Oliveira, P.; Fernandes, B.; Ferreira, E. C.; Alves, M. M.; Lopes, F.; Pereira, M. A.; Novais, P. A Review of Computational Modeling in Wastewater Treatment Processes. *ACS EST Water* **2024**, *4* (3), 784–804. https://doi.org/10.1021/acsestwater.3c00117.

(49) Eaton, A. D.; Clesceri, L. S.; Franson, M. A. H.; Rice, E. W.; Greenberg, A. E. Standard Methods for the Examination of Water and Wastewater, 2005.

(50) Kanehisa, M.; Sato, Y.; Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **2016**, *428,4*, 726–731.

(51) Torregrossa, D.; Leopold, U.; Hernández-Sancho, F.; Hansen, J. Machine Learning for Energy Cost Modelling in Wastewater Treatment Plants. *J Env. Manage* **2018**, *223*, 1061–1067.

(52) Asadi, A.; Verma, A.; Yang, K.; Mejabi, B. Wastewater Treatment Aeration Process Optimization: A Data Mining Approach. *J Env. Manage* **2017**, *203*, 630–639.

(53) Drainage Services Department. Data Collected Directly from the Government Department in Hong Kong S.A.R, 2017.

(54) Newhart, K. B.; Holloway, R. W.; Hering, A. S.; Cath, T. Y. Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Res.* **2019**, *157*, 498–513. https://doi.org/10.1016/j.watres.2019.03.030.

(55) Chaoui, A.; Rebija, K.; Chkaiti, K.; Laaouan, M.; Bourziza, R.; Sebari, K.; Elkhoumsi, W. Applications of Missing Data Imputation Methods in Wastewater Treatment Plants A Systematic Literature Review Using the Kitchenham Method. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 461. https://doi.org/10.14569/IJACSA.2023.0141049.

(56) Tieleman, T.; Hinton, G. Lecture 6.5-Rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude. *COURSERA Neural Netw. Mach. Learn.* **2012**, *4*, 2.

(57) Kingma, D. P.; Ba, J. L. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*; 2015; pp 1–15.

(58) Breiman, L. *Random For.* **2001**, *45*, 5–32.

(59) Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman and Hall/CRC, 1984. https://doi.org/10.1201/9781315139470.

(60) Drucker, H.; Surges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. *Adv Neural Inf Process Syst* **1997**, *1*, 155–161.

(61) Cristianini, N.; Ricci, E. Support Vector Machines, 2008.

(62) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; M., P.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res* **2011**, *12*, 2825–2830.

(63) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard,

M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv March 16, 2016. https://doi.org/10.48550/arXiv.1603.04467.

(64) Keras 3: A New Multi-Backend Keras, 2023. https://github.com/keras-team/keras (accessed 2023-12-28).

(65) Asghari, V.; Leung, Y. F.; Hsu, S. C. Deep neural network based framework for complex correlations in engineering metrics. *Adv Eng Inform.* **2020**, *44*, 101058.

(66) Wang, R.; Asghari, V.; Hsu, S. C.; Lee, C. J.; Chen, J. H. Detecting Corporate Misconduct through Random Forest in China's Construction Industry. *J Clean Prod* **2020**, *268*, 122266.

(67) Corbalá-Robles, L.; Picioreanu, C.; Loosdrecht, M. C. M.; Pérez, J. Analysing the Effects of the Aeration Pattern and Residual Ammonium Concentration in a Partial Nitritation-Anammox Process. *Env. Technol* **2016**, *37*, 694–702.

(68) Skouteris, G.; Arnot, T. C.; Jraou, M.; Feki, F.; Sayadi, S. Modeling Energy Consumption in Membrane Bioreactors for Wastewater Treatment in North Africa. *Water Env. Res* **2014**, *86*, 232–244.

(69) Chen, Z.; Ren, N.; Wang, A.; Zhang, Z. P.; Shi, Y. A Novel Application of TPAD-MBR System to the Pilot Treatment of Chemical Synthesis-Based Pharmaceutical Wastewater. *Water Res* **2008**, *42*, 3385–3392.

(70) Urakami, T.; Sasaki, J.; Suzuki, K. I.; Komagata, K. Characterization and Description of Hyphomicrobium Denitrificans Sp. Nov. *Int J Syst Bacteriol* **1995**, *45*, 528–532.

(71) Flood, B. E.; Jones, D. S.; Bailey, J. V. Sedimenticola Thiotaurini Sp. Nov., a Sulfur-Oxidizing Bacterium Isolated from Salt Marsh Sediments, and Emended Descriptions of the Genus Sedimenticola and Sedimenticola Selenatireducens. *Int J Syst Evol Microbiol* **2015**, *65*, 2522–2530.

(72) Collado, L.; Levican, A.; Perez, J.; Figueras, M. J. Arcobacter Defluvii Sp. Nov., Isolated from Sewage Samples. *Int J Syst Evol Microbiol* **2011**, *61*, 2155–2161.

(73) Chen, Q.; Ni, J.; Ma, T.; Liu, T.; Zheng, M. Bioaugmentation Treatment of Municipal Wastewater with Heterotrophic-Aerobic Nitrogen Removal Bacteria in a Pilot-Scale SBR. *Bioresour Technol* **2015**, *183*, 25–32.

(74) Sorokin, D. Y.; Tourova, T. P.; Lysenko, A. M.; Muyzer, G. Diversity of Culturable Halophilic Sulfur-Oxidizing Bacteria in Hypersaline Habitats. *Microbiology* **2006**, *152*, 3013–3023.

(75) Baumann, B.; Snozzi, M.; Zehnder, A. J. B.; Meer, J. R. Dynamics of Denitrification Activity of Paracoccus Denitrificans in Continuous Culture during Aerobic-Anaerobic Changes. *J Bacteriol* **1996**, *178*, 4367–4374.

(76) Dou, Q.; Yang, J.; Peng, Y.; Zhang, L. Multipathway of Nitrogen Metabolism Revealed by Genome-Centered Metatranscriptomics from Pyrite-Guided Mixotrophic Partial Denitrification/Anammox Installations. *Environ. Sci. Technol.* **2023**, *57* (51), 21791–21800. https://doi.org/10.1021/acs.est.3c08192.

(77) Narasingarao, P.; Häggblom, M. M. Sedimenticola Selenatireducens, Gen. Nov., Sp. Nov., an Anaerobic Selenate-Respiring Bacterium Isolated from Estuarine Sediment. *Syst Appl Microbiol* **2006**, *29*, 382–388.

(78) Altamia, M. A.; Shipway, J. R.; Concepcion, G. P.; Haygood, M. G.; Distel, D. L. Thiosocius Teredinicola Gen. Nov., Sp. Nov., a Sulfur-Oxidizing Chemolithoautotrophic Endosymbiont Cultivated from the Gills of the Giant Shipworm, Kuphus Polythalamius. *Int. J. Syst. Evol. Microbiol.* **2019**, *69* (3), 638–644. https://doi.org/10.1099/ijsem.0.003143.

(79) Vick, S. H. W.; Tetu, S. G.; Sherwood, N.; Pinetown, K.; Sestak, S.; Vallotton, P.; Elbourne, L. D. H.; Greenfield, P.; Johnson, E.; Barton, D.; Midgley, D. J.; Paulsen, I. T. Revealing Colonisation and Biofilm Formation of an Adherent Coal Seam Associated Microbial Community on a Coal Surface. *Int J Coal Geol* **2016**, *160–161*, 42–50.

(80) Çinar, Ö.; Hasar, H.; Kinaci, C. Modeling of Submerged Membrane Bioreactor Treating Cheese Whey Wastewater by Artificial Neural Network. *J Biotechnol* **2006**, *123*, 204–209.

(81) Heylen, K.; Vos, P.; Vekeman, B. Draft Genome Sequences of Eight Obligate Methane Oxidizers Occupying Distinct Niches Based on Their Nitrogen Metabolism. *Genome Announc* **2016**, *4*, 16–17.

(82) Ogiso, T.; Ueno, C.; Dianou, D.; Huy, T.; Katayama, A.; Kimura, M.; Asakawa, S. Methylomonas Koyamae Sp. Nov., a Type I Methane-Oxidizing Bacterium from Floodwater of a Rice Paddy Field. *Int J Syst Evol Microbiol* **2012**, *62*, 1832–1837.

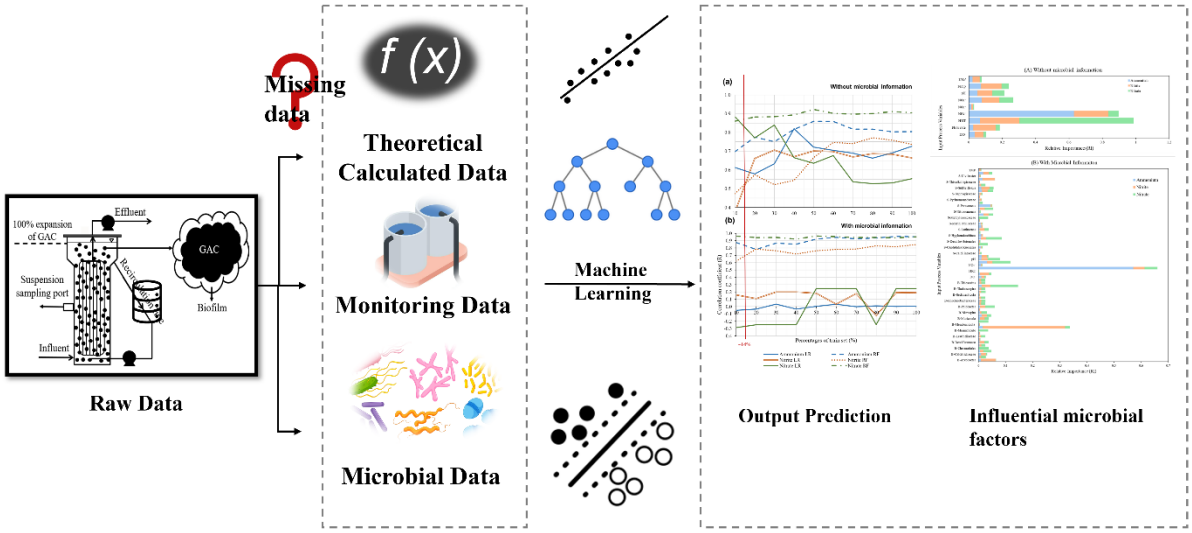
(83) Liesack, W.; Finster, K. Phylogenetic Analysis of Five Strains of Gram-Negative, Obligately Anaerobic, Sulfur-Reducing Bacteria and Description of Desulfuromusa Gen. Nov., Including Desulfuromusa Kysingii Sp. Nov., Desulfuromusa Bakii Sp. Nov., and Desulfuromusa Succinoxidans Sp. *Int J Syst Bacteriol* **1994**, *44*, 753–758.

(84) Xiong, W.; Ye, Y.; He, D.; He, S.; Xiang, Y.; Xiao, J.; Feng, W.; Wu, M.; Yang, Z.; Wang, D. Deregulation of Ribosome Biogenesis in Nitrite-Oxidizing Bacteria Leads to Nitrite Accumulation. *Environ. Sci. Technol.* **2023**, *57* (43), 16673–16684. https://doi.org/10.1021/acs.est.3c06002.

(85) Jung, Y. T.; Kim, J. H.; Oh, T. K.; Yoon, J. H. Mariniflexile Aquimaris Sp. Nov., Isolated from Seawater, and Emended Description of the Genus Mariniflexile Nedashkovskaya et Al. *Int J Syst Evol Microbiol* **2012**, *62*, 539–544.

(86) Yang, J.; Chen, Z.; Wang, X.; Zhang, Y.; Li, J.; Zhou, S. Elucidating Nitrogen Removal Performance and Response Mechanisms of Anammox under Heavy Metal Stress Using Big Data Analysis and Machine Learning. *Bioresour. Technol.* **2023**, *382*, 129143. https://doi.org/10.1016/j.biortech.2023.129143.

814 TOC Graphic



815